21 世纪全国本科院校电气信息类创新型应用人才培养规划教材

信息与通信专业英语

主　编　刘小佳



内 容 简 介

本书由课文、阅读材料和摘要阅读组成，内容分别选自各领域经典英文教材、国际顶级期刊及会议的论文摘要，紧扣信息、通信与计算三大主题，内容全面详实。

本书可作为信息与通信工程、信息安全、电子科学与技术、电子信息工程、计算机科学与技术和网络工程等专业本科生、研究生“专业英语”课程的教材，也可供相关专业技术人员学习和参考。

图书在版编目(CIP)数据

信息与通信专业英语/刘小佳主编．—北京：北京大学出版社，2015.4
（21世纪全国本科院校电气信息类创新型应用人才培养规划教材）
ISBN 978-7-301-25506-3

Ⅰ.①信…　Ⅱ.①刘…　Ⅲ.①信息技术—英语—高等学校—教材②通信工程—英语—高等学校—教材　Ⅳ.①H31

中国版本图书馆CIP数据核字（2015）第031991号

书　　名　信息与通信专业英语
著作责任者　刘小佳　主编
责 任 编 辑　程志强
标 准 书 号　ISBN 978-7-301-25506-3
出 版 发 行　北京大学出版社
地　　址　北京市海淀区成府路205号　100871
网　　址　http://www.pup.cn　　新浪微博：@北京大学出版社
电 子 信 箱　pup_6@163.com
电　　话　邮购部62752015　发行部62750672　编辑部62750667
印 刷 者
经 销 者　新华书店
787毫米×1092毫米　16开本　13.75印张　312千字
2015年4月第1版　2015年4月第1次印刷
定　　价　29.00元

前　言

多年来，专业英语教学一直想传达本科教育的国际化视野之理念，但收效甚微。许多专业英语教材尽管选择了非常地道的英文原文，但其内容止于“科普”，未能达到“专业”。事实上，文史哲中所推崇的原典阅读是一个极好的思路，本书正是在此基础上展开了有益的尝试。

首先，要能阅读较为专业的英语文本。作者从电子信息类与通信类专业领域经典英文教材入手，精选若干原文，模拟学生学习该专业课程原文教材的场景，以此提高专业英语的针对性。显然，这种直接阅读大师原典的做法不但能提高使用专业英语的能力，还能在专业方面打开知识的大门，为进一步阅读经典教材全文打下基础。

其次，专业英语写作也是很重要的内容。作者在每个领域选出了若干国际顶级期刊与会议的论文摘要，针对国内学生摘要写作较差的特点进行强化训练，以此培养他们科技论文写作的能力，并以摘要为起点抛砖引玉，令有能力的学生进行完整的英语科技论文写作。

在内容选取上，本书充分考虑了通信工程、信息安全、电子科学与技术、电子信息工程、计算机科学与技术和网络工程等专业的特色，紧扣信息、通信与计算三大主题，突出这三个领域的交叉与融合，不但为更多专业的学生提供了学习的内容，还注重培养各专业对其相关领域的关注了解，从而更好地提高学生的综合实力。

本书共分为16章，每章由一篇课文、一篇阅读材料和三篇摘要阅读组成。读者在学习完课文之后，应精心理解阅读材料。对于学有余力的学生，还应该寻根溯源去广泛阅读对应的国外教材，这样才能达到熟练掌握该领域英语的功效。为了帮助读者学习，作者对课文中的词汇给出了在该领域的准确释义，并精选课文中的阅读难点予以注释。需要指出的是，读者应仔细研读摘要的写法，不但能将英文翻译为中文，还能将已翻译的中文回译成英文，多年的教学实践证明，这种对比参照的双向翻译能收到良好的效果。

感谢西安邮电大学外国语学院袁小陆教授仔细审阅了全部书稿，并提出了许多中肯的建议！感谢西安邮电大学外国语学院诸多同事在教材编写中的鼎力相助！本书摘选诸多国外经典教材的英文文本，并遴选多篇顶级期刊会议的论文摘要，在此一并对原作者致谢！

由于编者学识所限，书中疏漏之处在所难免，恳请读者不吝赐教！

编　者

2014年12月

目　　录



UNIT 1

INFORMATION THEORY

Text: Introduction to Information Theory

Information theory answers two fundamental questions in communication theory: What is the ultimate data compression (answer: the entropy H), and what is the ultimate transmission rate of communication (answer: the channel capacity C). For this reason some consider information theory to be a subset of communication theory. We argue that it is much more. Indeed, it has fundamental contributions to make in statistical physics (thermodynamics), computer science (Kolmogorov complexity or algorithmic complexity), statistical inference (Occam's Razor: "The simplest explanation is best"), and to probability and statistics (error exponents for optimal hypothesis testing and estimation).

This chapter goes backward and forward through information theory and its naturally related ideas. Information theory intersects physics (statistical mechanics), mathematics (probability theory), electrical engineering (communication theory), and computer science (algorithmic complexity). We now describe the areas of intersection in greater detail.

Electrical Engineering (Communication Theory). In the early 1940s it was thought to be impossible to send information at a positive rate with negligible probability of error. Shannon surprised the communication theory community by proving that the probability of error could be made nearly zero for all communication rates below channel capacity.

The capacity can be computed simply from the noise characteristics of the channel[1]. Shannon further argued that random processes such as music and speech have an irreducible complexity below which the signal cannot be compressed. This he named the entropy, in deference to the parallel use of this word in thermodynamics, and argued that if the entropy of the source is less than the capacity of the channel, asymptotically error-free communication can be achieved.

Information theory today represents the extreme points of the set of all possible communication schemes. The data compression minimum $I(X;\hat{X})$ lies at one extreme of the set of communication ideas. All data compression schemes require description rates at least equal to this minimum. At the other extreme is the data transmission maximum $I(X;Y)$, known as the channel capacity. Thus, all modulation schemes and data compression schemes lie between these limits.

Information theory also suggests means of achieving these ultimate limits of communication. However, these theoretically optimal communication schemes, beautiful as they are, may turn out to be computationally impractical. It is only because of the computational feasibility of simple modulation and demodulation schemes that we use them rather than the random coding and nearest-neighbor decoding rule suggested byShannon's proof of the channel capacity theorem. Progress in integrated circuits and code design has enabled us to reap some of the gains suggested by Shannon's theory. Computational practicality was finally achieved by the advent of turbo codes. A good example of an application of the ideas of information theory is the use of error-correcting codes on compact discs and DVDs.

Recent work on the communication aspects of information theory has concentrated on network information theory: the theory of the simultaneous rates of communication from many senders to many receivers in the presence of interference and noise. Some of the trade-offs of rates between senders and receivers are unexpected, and all have a certain mathematical simplicity. A unifying theory, however, remains to be found.

Computer Science (Kolmogorov Complexity). Kolmogorov, Chaitin, and Solomonoff put forth the idea that the complexity of a string of data can be defined by the length of the shortest binary computer program for computing the string. Thus, the complexity is the minimal description length. This definition of complexity turns out to be universal, that is, computer independent, and is of fundamental importance[2]. Thus, Kolmogorov complexity lays the foundation for the theory of descriptive complexity. Gratifyingly, the Kolmogorov complexity K is approximately equal to the Shannon entropy H if the sequence is drawn at random from a distribution that has entropy H. So the tie-in between information theory and Kolmogorov complexity is perfect. Indeed, we consider Kolmogorov complexity to be more fundamental than Shannon entropy. It is the ultimate data compression and leads to a logically consistent procedure for inference.

There is a pleasing complementary relationship between algorithmic complexity and computational complexity. One can think about computational complexity (time complexity) and Kolmogorov complexity (program length or descriptive complexity) as two axes corresponding to program running time and program length. Kolmogorov complexity focuses on minimizing along the second axis, and computational complexity focuses on minimi-



zing along the first axis. Little work has been done on the simultaneous minimization of the two.

Physics(Thermodynamics). Statistical mechanics is the birthplace of entropy and the second law of thermodynamics. Entropy always increases. Among other things, the second law allows one to dismiss any claims to perpetual motion machines.

Mathematics(Probability Theory and Statistics). The fundamental quantities of information theory—entropy, relative entropy, and mutual information—are defined as functionals of probability distributions[3]. In turn, they characterize the behavior of long sequences of random variables and allow us to estimate the probabilities of rare events (large deviation theory) and to find the best error exponent in hypothesis tests.

Philosophy of Science (Occam's Razor). William of Occam said "Causes shall not be multiplied beyond necessity," or to paraphrase it, "The simplest explanation is best." [4] Solomonoff and Chaitin argued persuasively that one gets a universally good prediction procedure if one takes a weighted combination of all programs that explain the data and observes what they print next. Moreover, this inference will work in many problems not handled by statistics. For example, this procedure will eventually predict the subsequent digits of π. When this procedure is applied to coin flips that come up heads with probability 0. 7, this too will be inferred. When applied to the stock market, the procedure should essentially find all the "laws" of the stock market and extrapolate them optimally. In principle, such a procedure would have found Newton's laws of physics. Of course, such inference is highly impractical, because weeding out all computer programs that fail to generate existing data will take impossibly long. We would predict what happens tomorrow a hundred years from now.

Economics (Investment). Repeated investment in a stationary stock market results in an exponential growth of wealth. The growth rate of the wealth is a dual of the entropy rate of the stock market. The parallels between the theory of optimal investment in the stock market and information theory are striking. We develop the theory of investment to explore this duality.

Computation vs. Communication. As we build larger computers out of smaller components, we encounter both a computation limit and a communication limit. Computation is communication limited and communication is computation limited. These become intertwined, and thus all of the developments in communication theory via information theory should have a direct impact on the theory of computation.

New Words and Phrases

transmission *n.* 传输
subset *n.* 子集
thermodynamics *n.* 热力学
algorithmic *adj.* 算法的
statistical inference 统计推断
exponent *n.* 指数
optimal *adj.* 最佳的，最优的
hypothesis *n.* 假设
intersect *vi.* 交叉，相交
complexity *n.* 复杂度
universal *adj.* 通用的，广泛的，普遍的，宇宙的
intersection *n.* 交集
mutual *adj.* 相互的，共同的
negligible *adj.* 微小的；可忽略的
irreducible *adj.* 不能分解的
entropy *n.* 熵
probability *n.* 概率
deference *n.* 顺从；尊重
asymptotically *adv.* 渐近地
modulation *n.* 调制
feasibility *n.* 可行性
integrated circuit 集成电路
reap *vt.* & *vi.* 收获
turbo code turbo 编码
simultaneous *adj.* 同时的，同步的
trade-off *n.* 权衡
unifying *vt.* 使统一
binary *adj.* 二进制的，二元的
tie-in *n.* 接头
compression *n.* 压缩
complementary *adj.* 互补的，补充的
axis *n.* 轴，轴线
perpetual motion 永恒运动
deviation *n.* 偏差；偏向
subsequent *adj.* 后来的，随后的
extrapolate *vt.* & *vi.* 推断，推算
optimally *adj.* 最佳
weed out 淘汰，剔除
exponential *n.* 指数
dual *n.* 对偶
    adj. 指数的
intertwine *vi.* & *vt.* 交织，纠缠
theorem *n.* 定理

Notes

1. The capacity can be computed simply from the noise characteristics of the channel.
（信道）容量只用信道的噪声特征即可计算。

2. Thus, the complexity is the minimal description length. This definition of complexity turns out to be universal, that is, computer independent, and is of fundamental importance.

因此，复杂度是最小描述长度。这个复杂度定义是通用的，也就是说，它与具体计算机无关并且有着根本的重要性。



3. The fundamental quantities of information theory—entropy, relative entropy, and mutual information—are defined as functionals of probability distributions.

信息论的基本量——熵、相对熵、互信息——被定义为与概率分布有关的函数。

4. (Occam's Razor) "Causes shall not be multiplied beyond necessity," or to paraphrase it, "The simplest explanation is best."

"如无必要，勿增缘由"，也就是"简单的解释就是最好的"。

Occam's Razor 奥卡姆剃刀原理：如果你有两个原理，它们都能解释观测到的事实，那么你应该使用简单的那个，直到发现更多的证据。对于现象最简单的解释往往比复杂的解释更正确。如果你有两个类似的解决方案，选择最简单的。需要最少假设的解释最有可能是正确的。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. error-correcting
2. probability theory
3. algorithmic complexity
4. large deviation theory
5. random processes
6. modulation schemes
7. statistical inference
8. negligible probability

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

theorem	compression	elements	length
capacity	entropy	random	probability
codes	parallels	transmission	mutual
defined	block	output	exponent
continuous	duality	redundancy	constructing
satisfy	ratio	capacity	concept

1. The relative ________ D arises as the ________ in the probability of error in a hypothesis test between two distributions. It is a natural measure of distance between distributions.



2. There are a number of ________ between information theory and the theory of investment in a stock market. A stock market is ________ by a random vector X whose ________ are nonnegative numbers equal to the ________ of the price of a stock at the end of a day to the price at the beginning of the day.

3. Entropy is the uncertainty of a single ________ variable. We can define conditional entropy $H(X \mid Y)$, which is the entropy of a random variable conditional on the knowledge of another random variable.

4. We now define the (information) ________ of the channel as the maximum of the ________ information between the input and ________ over all distributions on the input that ________ the power constraint.

5. We now introduce the concept of differential entropy, which is the entropy of a ________ random variable. Differential entropy is also related to the shortest description ________ and is similar in many ways to the entropy of a discrete random variable. But there are some important differences, and there is need for some care in using the ________.

6. The channel coding ________ promises the existence of block codes that will allow us to transmit information at rates below ________ with an arbitrarily small ________ of error if the block length is large enough.

7. Although the theorem shows that there exist good ________ with arbitrarily small probability of error for long ________ lengths, it does not provide a way of ________ the best codes.

8. There is a ________ between the problems of data ________ and data transmission. During compression, we remove all the ________ in the data to form the most compressed version possible, whereas during data ________, we add redundancy in a controlled fashion to combat errors in the channel.

Ⅲ. Translate the following paragraphs into Chinese.

1. At first sight, information theory and gambling seem to be unrelated. But as we shall see, there is strong duality between the growth rate of investment in a horse race and the entropy rate of the horse race. Indeed, the sum of the growth rate and the entropy rate is a constant. In the process of proving this, we shall argue that the financial value of side information is equal to the mutual information between the horse race and the side information. The horse race is a special case of investment in the stock market. We also show how to use a pair of identical gamblers to compress a sequence of random variables by an amount equal to the growth rate of wealth on that sequence. Finally, we use these gambling techniques to estimate the entropy rate of English.

2. This enables us to divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences. Most of our attention will be on the typical sequences. Any property that is proved for the typical sequences will then be true with high probability and will determine the average behavior of a large sample.

Reading: Channel Capacity

What do we mean when we say that A communicates with B? We mean that the physical acts of A have induced a desired physical state in B. This transfer of information is a physical process and therefore is subject to the uncontrollable ambient noise and imperfections of the physical signaling process itself. The communication is successful if the receiver B and the transmitter A agree on what was sent.

We find the maximum number of distinguishable signals for n uses of a communication channel. This number grows exponentially with n, and the exponent is known as the



channel capacity. The characterization of the channel capacity (the logarithm of the number of distinguishable signals) as the maximum mutual information is the central and most famous success of information theory.

The mathematical analog of a physical signaling system is shown in Figure 1.1. Source symbols from some finite alphabet are mapped into some sequence of channel symbols, which then produces the output sequence of the channel. The output sequence is random but has a distribution that depends on the input sequence. From the output sequence, we attempt to recover the transmitted message.

Each of the possible input sequences induces a probability distribution on the output sequences. Since two different input sequences may give rise to the same output sequence, the inputs are confusable. In the next few sections, we show that we can choose a "nonconfusable" subset of input sequences so that with high probability there is only one highly likely input that could have caused the particular output. We can then reconstruct the input sequences at the output with a negligible probability of error. By mapping the source into the appropriate "widely spaced" input sequences to the channel, we can transmit a message with very low probability of error and reconstruct the source message at the output. The maximum rate at which this can be done is called the capacity of the channel.

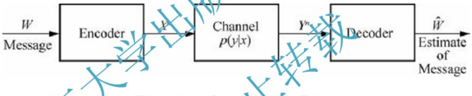


Figure 1.1 Communication System

Definition We define *a discrete channel* to be a system consisting of an input alphabet X and output alphabet Y and a probability transition matrix $p(y \mid x)$ that expresses the probability of observing the output symbol y given that we send the symbol x. The channel is said to be *memoryless* if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.

Definition We define the "*information*" *channel capacity* of a discrete memoryless channel as where the maximum is taken over all possible input distributions $p(x)$.

$$C = \max_{p(x)} I(X;Y) \tag{1-1}$$

We shall soon give an operational definition of channel capacity as the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error. Shannon's second theorem establishes that the information channel capacity is equal to the operational channel capacity. Thus, we drop the word information in most discussions of channel capacity.

There is a duality between the problems of data compression and data transmission. During compression, we remove all the redundancy in the data to form the most com-



pressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel. Later we show that a general communication system can be broken into two parts and that the problems of data compression and data transmission can be considered separately.

New Words and Phrases

ambient *adj*. 背景的
imperfection *n*. 缺陷
exponentially *adv*. 成指数地
exponent *n*. 指数，幂
analog *n*. 模拟
negligible *adj*. 可忽略不计的
discrete *adj*. 离散，离散的
arbitrarily *adv*. 任意地
duality *n*. 对偶，二元性
compression *n*. 压缩
redundancy *n*. 冗余，冗余度

Exercises

Ⅰ. Answer the following questions.

1. What is the discrete channel?
2. How to calculate the channel capacity?
3. What is theredundancy?

Ⅱ. Translate the following sentences into Chinese.

1. This transfer of information is a physical process and therefore is subject to the uncontrollable ambient noise and imperfections of the physical signaling process itself.

____

____

2. Source symbols from some finite alphabet are mapped into some sequence of channel symbols, which then produces the output sequence of the channel.

____

____

3. By mapping the source into the appropriate "widely spaced" input sequences to the channel, we can transmit a message with very low probability of error and reconstruct the source message at the output.

____

____

4. During compression, we remove all the redundancy in the data to form the most compressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel.

____

____

____

Abstract Reading

On the Capacity of the Two-user Gaussian Causal Cognitive Interference Channel

This paper considers the two-user Gaussian causal cognitive interference channel (GCCIC), which consists of two source-destination pairs that share the same channel and where one full-duplex cognitive source can causally learn the message of the primary source through a noisy link. The GCCIC is an interference channel with unilateral source cooperation that better models practical cognitive radio networks than the commonly used model which assumes that one source has perfect noncausal knowledge of the other source's message. First, the sum-capacity of the symmetric GCCIC is determined to within a constant gap. Then, the insights gained from the study of the symmetric GCCIC are extended to more general cases. In particular, the whole capacity region of the Gaussian Z-channel, i. e., when there is no interference from the primary user, and of the Gaussian S-channel, i. e., when there is no interference from the secondary user, are both characterized to within 2 bits. The fully connected general, i. e., no-symmetric, GCCIC is also considered and its capacity region is characterized to within 2 bits when, roughly speaking, the interference is not weak at both receivers. The parameter regimes where the GCCIC is equivalent, in terms of generalized degrees-of-freedom, to the noncooperative interference channel (i. e., unilateral causal cooperation is not useful), to the non-causal cognitive interference channel (i. e., causal cooperation attains the ultimate limit of cognitive radio technology), and to bilateral source cooperation are identified. These comparisons shed light into the parameter regimes and network topologies that in practice might provide an unbounded throughput gain compared to currently available (non cognitive) technologies.

Channel Coding and Lossy Source Coding Using a Generator of Constrained Random Numbers

Stochastic encoders for channel coding and lossy source coding are introduced with a

rate close to the fundamental limits, where the only restriction is that the channel input alphabet and the reproduction alphabet of the lossy source code are finite. Random numbers, which satisfy a condition specified by a function and its value, are used to construct stochastic encoders. The proof of the theorems is based on the hash property of an ensemble of functions, where the results are extended to general channels/sources and alternative formulas are introduced for channel capacity and the rate-distortion region. Since an ensemble of sparse matrices has a hash property, we can construct a code by using sparse matrices.

Is Non-unique Decoding Necessary?

In multiterminal communication systems, signals carrying messages meant for different destinations are often observed together at any given destination receiver. Han and Kobayashi proposed a receiving strategy, which performs a joint unique decoding of messages of interest along with a subset of messages, which are not of interest. It is now well-known that this provides an achievable region, which is, in general, larger than if the receiver treats all messages not of interest as noise. Nair and El Gamal and Chong, Motani, Garg, and El Gamal independently proposed a generalization called indirect or nonunique decoding where the receiver uses the codebook structure of the messages to uniquely decode only its messages of interest. Nonunique decoding has since been used in various scenarios. The main result in this paper is to provide an interpretation and a systematic proof technique for why nonunique decoding, in all known cases where it has been employed, can be replaced by a particularly designed joint unique decoding strategy, without any penalty from a rate region viewpoint.

UNIT 2

ALGORITHMS

Text: Analyzing Algorithms

Analyzing an algorithm has come to mean predicting the resources that the algorithm requires. Occasionally, resources such as memory, communication bandwidth, or computer hardware are of primary concern, but most often it is computational time that we want to measure. Generally, by analyzing several candidate algorithms for a problem, we can identify a most efficient one. Such analysis may indicate more than one viable candidate, but we can often discard several inferior algorithms in the process.

Before we can analyze an algorithm, we must have a model of the implementation technology that we will use, including a model for the resources of that technology and their costs. We shall assume a generic one processor, random-access machine (RAM)[1] model of computation as our implementation technology and understand that our algorithms will be implemented as computer programs. In the RAM model, instructions are executed one after another, with no concurrent operations.

Strictly speaking, we should precisely define the instructions of the RAM model and their costs. To do so, however, would be tedious and would yield little insight into algorithm design and analysis. Yet we must be careful not to abuse the RAM model. For example, what if a RAM had an instruction that sorts? Then we could sort in just one instruction. Such a RAM would be unrealistic, since real computers do not have such instructions. Our guide, therefore, is how real computers are designed. The RAM model contains instructions commonly found in real computers: arithmetic (such as add, subtract, multiply, divide, remainder, floor, ceiling), data movement (load, store, copy), and control (conditional and unconditional branch, subroutine call and return)[2]. Each such instruction takes a constant amount of time.



The data types in the RAM model are integer and floating point (for storing real numbers). Although we typically do not concern ourselves with precision in this book, in some applications precision is crucial. We also assume a limit on the size of each word of data. For example, when working with inputs of size n, we typically assume that integers are represented by $c \lg n$ bits for some constant $c \geqslant 1$. We require $c \geqslant 1$ so that each word can hold the value of n, enabling us to index the individual input elements, and we restrict c to be a constant so that the word size does not grow arbitrarily. (If the word size could grow arbitrarily, we could store huge amounts of data in one word and operate on it all in constant time—clearly an unrealistic scenario.)

Real computers contain instructions not listed above, and such instructions represent a gray area in the RAM model. For example, isexponentiation a constant time instruction? In the general case, no; it takes several instructions to compute x^y when x and y are real numbers. In restricted situations, however, exponentiation is a constant-time operation. Many computers have a "shift left" instruction, which in constant time shifts the bits of an integer by k positions to the left. In most computers, shifting the bits of an integer by one position to the left is equivalent to multiplication by 2, so that shifting the bits by k positions to the left is equivalent to multiplication by 2^k. Therefore, such computers can compute $2k$ in one constant-time instruction by shifting the integer 1 by k positions to the left, as long as k is no more than the number of bits in a computer word. We will endeavor to avoid such gray areas in the RAM model, but we will treat computation of 2^k as a constant-time operation when k is a small enough positive integer.

In the RAM model, we do not attempt to model the memoryhierarchy that is common in contemporary computers. That is, we do not model caches or virtual memory. Several computational models attempt to account for memory-hierarchy effects, which are sometimes significant in real programs on real machines. A handful of problems in this book examine memory-hierarchy effects, but for the most part, the analyses in this book will not consider them. Models that include the memory hierarchy are quite a bit more complex than the RAM model, and so they can be difficult to work with. Moreover, RAM-model analyses are usually excellent predictors of performance on actual machines.

Analyzing even a simple algorithm in the RAM model can be a challenge. The mathematical tools required may includecombinatorics, probability theory, algebraic dexterity, and the ability to identify the most significant terms in a formula[3]. Because the behavior of an algorithm may be different for each possible input, we need a means for summarizing that behavior in simple, easily understood formulas.

Even though we typically select only one machine model to analyze a given algorithm, we still face many choices in deciding how to express our analysis. We would like a way

that is simple to write and manipulate, shows the important characteristics of an algorithm's resource requirements, and suppresses tedious details.

Analysis of Insertion Sort

The time taken by the INSERTION-SORT procedure depends on the input: sorting a thousand numbers takes longer than sorting three numbers. Moreover, INSERTION-SORT can take different amounts of time to sort two input sequences of the same size depending on how nearly sorted they already are. In general, the time taken by an algorithm grows with the size of the input, so it is traditional to describe the running time of a program as a function of the size of its input. To do so, we need to define the terms "running time" and "size of input" more carefully.

The best notion for input size depends on the problem being studied. For many problems, such as sorting or computing discrete Fourier transforms, the most natural measure is the number of items in the input—for example, the array size n for sorting. For many other problems, such as multiplying two integers, the best measure of input size is the total number of bits needed to represent the input in ordinary binary notation. Sometimes, it is more appropriate to describe the size of the input with two numbers rather than one. For instance, if the input to an algorithm is a graph, the input size can be described by the numbers of vertices and edges in the graph. We shall indicate which input size measure is being used with each problem we study.

The running time of an algorithm on a particular input is the number of primitive operations or "steps" executed. It is convenient to define the notion of step so that it is as machine-independent as possible. For the moment, let us adopt the following view. A constant amount of time is required to execute each line of our pseudocode. One line may take a different amount of time than another line, but we shall assume that each execution of the ith line takes time c_i, where c_i is a constant. This viewpoint is in keeping with the RAM model, and it also reflects how the pseudocode would be implemented on most actual computers.

New Words and Phrases

bandwidth *n.* 带宽
inferior *adj.* 下等的
　　n. 下级，次品
implementation *n.* 实现，履行，安装启用
generic *adj.* 泛化的
execute *vt.* 执行，运行
concurrent *adj.* 同时发生的；同时完成的；同时存在的
　　n. 共点；同时发生的事件
remainder *n.* 余数
　　adj. 剩余的
ceiling *n.* 向上取整



subroutine *n.* 子程序
integer *n.* 整数
floating *n.* 浮点数
scenario *n.* 场景
exponentiation *n.* 求幂，取幂
equivalent *adj.* 相等的，相当的，等效的；等价的，等积的
n. 对等物
endeavor *vt.* & *vi.* 尝试；尽力
n. 努力，尽力
hierarchy *n.* 分层，层次
cache *n.* 高速缓冲存储器
memory-hierarchy *n.* 存储层次；分级存储器体系
predictor *n.* 预测器，预测程序
combinatoric *adj.* 组合学的
manipulate *vt.* 处理，控制
insertion-sort *n.* 插入排序
discrete *adj.* 离散的
binay *adj.* 二元的；二进制的
n. 二进制数
vertices *n.* 顶点
primitive *n.* 元数据类型
preudocode *n.* 伪代码
messy formula *n.* 延迟公式
manipulate *vt.* 控制，操作
notation *n.* 记号，符号

Notes

1. random-access machine：随机存取存储器（RAM）又称作"随机存储器"，是与 CPU 直接交换数据的内部存储器，也叫主存。它可以随时读写，而且速度很快，通常作为操作系统或其他正在运行的程序的临时数据存储媒介。

2. The RAM model contains instructions commonly found in real computers：arithmetic (such as add，subtract，multiply，divide，remainder，floor，ceiling)，data movement (load，store，copy)，and control (conditional and unconditional branch，subroutine call and return). RAM 模型中包含的指令一般可以在计算机中找到：算术(如加、减、乘、除、余数、向下取整、向上取整)，数据传送(如加载、存储、复制)和控制(如有条件转移和无条件转移、子程序调用和返回)。

3. The mathematical tools required may includecombinatorics，probability theory，algebraic dexterity，and the ability to identify the most significant terms in a formula. 所需的数学工具可能包括：组合数学、概率论、代数技巧及识别公式中有效项的能力。

Exercises

Ⅰ. Translate the following words and phrases into Chinese.

1. implementation technology

2. concurrent operation
3. unrealistic scenario
4. positive integer
5. virtual memory
6. running time
7. input size
8. discrete Fourier transforms
9. binary notation
10. primitive operation

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

convenient	implementation	assume	introduces
distribution	software	minimum	reverse
averaging	notion	scenario	alternatives
random	input	infinitely	insertion
probabilities	theoretical	array	efficient

1. If computers were ________ fast, any correct method for solving a problem would do. You would probably want your ________ to be within the bounds of good ________ engineering practice (for example, your implementation should be well designed and documented), but you would most often use whichever method was the easiest to implement.

2. In order to analyze many algorithms, including the hiring problem, we use indicator ________ variables. Indicator random variables provide a ________ method for converting between ________ and expectations.

3. We must be very careful in deciding on the distribution of ________. For some problems, we may reasonably ________ something about the set of all possible inputs, and then we can use probabilistic analysis as a technique for designing an ________ algorithm and as a means for gaining insight into a problem. For other problems, we cannot describe a reasonable input ________, and in these cases we cannot use probabilistic analysis.

4. This ________ serves as a model for a common computational paradigm. We often need to find the maximum or ________ value in a sequence by examining each element of the sequence and maintaining a current "winner". The hiring problem models how often we update our ________ of which element is currently winning.

5. Most of the algorithms we discuss have great practical utility. We therefore address implementation concerns and other engineering issues. We often provide practical

________ to the few algorithms that are primarily of ________ interest.

6. In some cases, we assume that the inputs conform to a known probability distribution, so that we are ________ the running time over all possible inputs.

7. Having specified the ________ sort algorithm, we then argue that it correctly sorts, and we analyze its running time. The analysis ________ a notation that focuses on how that time increases with the number of items to be sorted.

8. In our analysis of insertion sort, we looked at both the best case, in which the input ________ was already sorted, and the worst case, in which the input array was ________ sorted.

Ⅲ. Translate the following paragraphs into Chinese.

1. The worst-case running time of an algorithm gives us an upper bound on the running time for any input. Knowing it provides a guarantee that the algorithm will never take any longer. We need not make some educated guess about the running time and hope that it never gets much worse.

2. For some algorithms, the worst case occurs fairly often. For example, in searching a database for a particular piece of information, the searching algorithm's worst case will often occur when the information is not present in the database. In some applications, searches for absent information may be frequent.

3. Often, we shall assume that all inputs of a given size are equally likely. In practice, this assumption may be violated, but we can sometimes use a randomized algorithm, which makes random choices, to allow a probabilistic analysis and yield an expected running time.

Reading: Multithreaded Algorithms

The vast majority of algorithms are serial algorithms suitable for running on a uniprocessor computer in which only one instruction executes at a time. In this paper, we shall extend our algorithmic model to encompass parallel algorithms, which can run on a multiprocessor computer that permits multiple instructions to execute concurrently. In particular, we shall explore the elegant model of dynamic multithreaded algorithms, which are amenable to algorithmic design and analysis, as well as to efficient implementation in practice.

Parallel computers—computers with multiple processing units—have become increasingly common, and they span a wide range of prices and performance. Relatively inexpensive desktop and laptop chip multiprocessors contain a single multicore integrated-circuit chip that houses multiple processing "cores", each of which is a full-fledged processor that can access a common memory. At an intermediate price/performance point are clusters built from individual computers—often simple PC-class machines—with a dedicated network interconnecting them. The highest-priced machines are supercomputers, which often use a combination of custom architectures and custom networks to deliver the highest performance in terms of instructions executed per second. Multiprocessor computers have been around, in one form or another, for decades. Although the computing community settled on the random-access machine model for serial computing early on in the history of computer science, no single model for parallel computing has gained as wide acceptance. A major reason is that vendors have not agreed on a single architectural model for parallel computers. For example, some parallel computers feature shared memory, where each processor can directly access any location of memory. Other parallel computers employ distributed memory, where each processor's memory is private, and an explicit message must be sent between processors in order for one processor to access the memory of another. With the advent of multicore technology, however, every new laptop and desktop machine is now



a shared-memory parallel computer and the trend appears to be toward shared-memory multiprocessing. Although time will tell, that is the approach we shall take in this paper.

One common means of programming chip multiprocessors and other shared memory parallel computers is by using static threading, which provides a software abstraction of "virtual processors", or threads, sharing a common memory. Each thread maintains an associated program counter and can execute code independently of the other threads. The operating system loads a thread onto a processor for execution and switches it out when another thread needs to run. Although the operating system allows programmers to create and destroy threads, these operations are comparatively slow. Thus, for most applications, threads persist for the duration of a computation, which is why we call them "static". Unfortunately, programming a shared-memory parallel computer directly using static threads is difficult and error-prone. One reason is that dynamically partitioning the work among the threads so that each thread receives approximately the same load turns out to be a complicated undertaking. For any but the simplest of applications, the programmer must use complex communication protocols to implement a scheduler to load-balance the work. This state of affairs has led toward the creation of concurrency platforms, which provide a layer of software that coordinates, schedules, and manages the parallel-computing resources. Some concurrency platforms are built as runtime libraries, but others provide full-fledged parallel languages with compiler and runtime support.

Dynamic Multithreaded Programming

One important class of concurrency platform is dynamic multithreading, which is the model we shall adopt in this paper. Dynamic multithreading allows programmers to specify parallelism in applications without worrying about communication protocols, load balancing, and other vagaries of static-thread programming. The concurrency platform contains a scheduler, which load-balances the computation automatically, thereby greatly simplifying the programmer's chore. Although the functionality of dynamic-multithreading environments is still evolving, almost all support two features: nested parallelism and parallel loops. Nested parallelism allows a subroutine to be "spawned", allowing the caller to proceed while the spawned subroutine is computing its result. A parallel loop is like an ordinary for loop, except that the iterations of the loop can execute concurrently.

These two features form the basis of the model for dynamic multithreading that we shall study in this paper. A key aspect of this model is that the programmer needs to specify only the logical parallelism within a computation, and the threads within the underlying concurrency platform schedule and load-balance the computation among themselves. We shall investigate multithreaded algorithms written for this model, as well how the underly-

ing concurrency platform can schedule computations efficiently.

Our model for dynamic multithreading offers several important advantages.

(1) It is a simple extension of our serial programming model. We can describe a multithreaded algorithm by adding to our pseudocode just three "concurrency" keywords: parallel, spawn, and sync. Moreover, if we delete these concurrency keywords from the multithreaded pseudocode, the resulting text is serial pseudocode for the same problem, which we call the "serialization" of the multithreaded algorithm.

(2) It provides a theoretically clean way to quantify parallelism based on the notions of "work" and "span".

(3) Many multithreaded algorithms involving nested parallelism follow naturally from the divide-and-conquer paradigm. Moreover, just as serial divide-and-conquer algorithms lend themselves to analysis by solving recurrences, so do multithreaded algorithms.

(4) The model is faithful to how parallel-computing practice is evolving. A growing number of concurrency platforms support one variant or another of dynamic multithreading, including Cilk, Cilk++, Open MP, Task Parallel Library, and Threading Building Blocks.

New Words and Phrases

multithreaded algorithms 多线程算法
uniprocessor *n.* 单机处理
encompass *vt.* 围绕，包围
multiprocessor *n.* 多处理理器
parallel computer 并行计算机
span *vt.* 跨越，测量
multicore *n.* 多核
circuit chip 电路芯片
full-fledged *adj.* 成熟的；完全的；完备的
cluster *n.* 群集
vendor *n.* 经销商
explicit *adj.* 明确的，清楚的，详述的
static *adj.* 静止的，静电的
abstraction *n.* 抽象
switch *n.* 开关；转换，转换器
  vt. & *vi.* 转变，改变，转换，关闭电流
  vt. 转换，迅速转动
  vi. 交换，调换
error-prone *adj.* 易错的
partition *n.* 划分，分割，隔离物
  vt. 区分，分割
protocol *n.* （数据传递的）协议；科学实验报告（或计划）
compiler *n.* 编译程序
dynamic multithreading 动态多线程
vagary *n.* 奇想，奇特行为
spawn *vt.* & *vi.* 大量产生，运行子程序
subroutine *n.* 子程序
loop *n.* 循环
pseudocode *n.* 伪代码
sync *n.* 同步
serialization *n.* 序列化
divide-and-conquer paradigm 分而治之范式

Exercises

Ⅰ. Answer the following questions.

1. What is parallel computer?

2. Why programming a shared-memory parallel computer directly using static threads is difficult and error-prone?

3. What does dynamic multithreading allow programmers to do?

4. What important advantages does our model for dynamic multithreading offer?

Ⅱ. Translate the following sentences into Chinese.

1. In particular, we shall explore the elegant model of dynamic multithreaded algorithms, which are amenable to algorithmic design and analysis, as well as to efficient implementation in practice.

____

____

____

____

____

2. Parallel computers—computers with multiple processing units—have become increasingly common, and they span a wide range of prices and performance.

____

____

____

3. One common means of programming chip multiprocessors and other shared memory parallel computers is by using static threading, which provides a software abstraction of "virtual processors", or threads, sharing a common memory.

____

____

____

____

____

4. Dynamic multithreading allows programmers to specify parallelism in applications without worrying about communication protocols, load balancing, and other vagaries of static-thread programming.

Abstract Reading

On the Complexity of Time-dependent Shortest Paths

We investigate the complexity of shortest paths in time-dependent graphs where the costs of edges (that is, edge travel times) vary as a function of time, and as a result the shortest path between two nodes s and d can change over time. Our main result is that when the edge cost functions are (polynomial-size) piecewise linear, the shortest path from s to d can change $n^{\Theta(\log n)}$ times, settling a several-year-old conjecture of Dean (Technical Reports, 1999, 2004). However, despite the fact that the arrival time function may have superpolynomial complexity, we show that a *minimum delay* path for any departure time interval can be computed in polynomial time. We also show that the complexity is polynomial if the slopes of the linear function come from a restricted class and describe an efficient scheme for computing a $(1+\varepsilon)$-approximation of the travel time function.

Algorithms for Partition of Some Class of Graphs under Compaction and Vertex-compaction

The compaction problem is to partition the vertices of an input graph G onto the vertices of a fixed target graph H, such that adjacent vertices of G remain adjacent in H, and every vertex and non-loop edge of H is covered by some vertex and edge of G respectively, i. e. , the partition is a homomorphism of G onto H (except the loop edges). Various computational complexity results, including both NP-completeness and polynomial time solvability, have been presented earlier for this problem for various classes of target graphs H. In this paper, we pay attention to the input graphs G, and present polynomial time algorithms for the problem for some class of input graphs, keeping the target graph H general as any reflexive or irreflexive graph. Our algorithms also give insight as for which instances of the input graphs, the problem could possibly be NP-complete for certain target graphs. With the help of our results, we are able to further refine the structure of the input graph that would be necessary for the problem to be possibly NP-complete, when the

target graph is a cycle. Thus, when the target graph is a cycle, we enhance the class of input graphs for which the problem is polynomial time solvable. We also present analogous results for a variation of the compaction problem, which we call the vertex-compaction problem. Using our results, we also provide important relationships between compaction, retraction, and vertex-compaction to cycles.

Sublinear Algorithms for Approximating String Compressibility

We raise the question of approximating the compressibility of a string with respect to a fixed compression scheme, in sublinear time. We study this question in detail for two popular lossless compression schemes: run-length encoding (RLE) and a variant of Lempel-Ziv (LZ77), and present sublinear algorithms for approximating compressibility with respect to both schemes. We also give several lower bounds that show that our algorithms for both schemes cannot be improved significantly.

Our investigation of LZ77 yields results whose interest goes beyond the initial questions we set out to study. In particular, we prove combinatorial structural lemmas that relate the compressibility of a string with respect to LZ77 to the number of distinct short substrings contained in it (its ℓth subword complexity, for small ℓ). In addition, we show that approximating the compressibility with respect to LZ77 is related to approximating the support size of a distribution.

UNIT 3

IMAGE COMPRESSION

Text: Approaches to Image Compression

An image compression method is normally designed for a specific type of image, and this section lists various approaches to compressing images of different types. Only the general principles are discussed here.

Approach 1: This is appropriate for bi-level images[1]. A pixel in such an image is represented by one bit. Applying the principle of image compression to a bi-level image therefore means that the immediate neighbors of a pixel P tend to be identical to P. Thus, it makes sense to use run-length encoding[2] (RLE) to compress such an image. A compression method for such an image may scan it in raster order (row by row) and compute the lengths of runs of black and white pixels. The lengths are encoded by variable-length codes and are written on the compressed stream. It should be stressed that this is just an approach to bi-level image compression. The details of specific methods vary. For instance, a method may scan the image column by column or in zigzag, it may convert the image to a quad tree, or it may scan it region by region using a space-filling curve[3].

Approach 2: Also for bi-level images. The principle of image compression tells us that the neighbors of a pixel tend to be similar to the pixel. We can extend this principle and conclude that if the current pixel has color c (where c is either black or white), then pixels of the same color seen in the past (and also those that will be found in the future) tend to have the same immediate neighbors.

This approach looks at n of the near neighbors of the current pixel and considers them an n-bit number. This number is the context of the pixel. In principle there can be 2^n contexts, but because of image redundancy we expect them to be distributed in a nonuniform way. Some contexts should be common while others will be rare.



The encoder counts how many times each context has already been found for a pixel of color c, and assigns probabilities to the contexts accordingly. If the current pixel has color c and its context has probability p, the encoder can use adaptive arithmetic coding to encode the pixel with that probability. This approach is used by JBIG[4].

Next, we turn to grayscale images. A pixel in such an image is represented by n bits and can have one of 2^n values. Applying the principle of image compression to a grayscale image implies that the immediate neighbors of a pixel P tend to be similar to P, but are not necessarily identical. Thus, RLE should not be used to compress such an image. Instead, two approaches are discussed.

Approach 3: Separate the grayscale image into n bi-level images and compress each with RLE and prefix codes. The principle of image compression seems to imply intuitively that two adjacent pixels that are similar in the grayscale image will be identical in most of the n bi-level images. This, however, is not true, as the following example makes clear. Imagine a grayscale image with $n=4$ (i. e. , 4-bit pixels, or 16 shades of gray). The image can be separated into four bi-level images. If two adjacent pixels in the original grayscale image have values 0000 and 0001, then they are similar. They are also identical in three of the four bi-level images. However, two adjacent pixels with values 0111 and 1000 are also similar in the grayscale image (their values are 7 and 8, respectively) but differ in all four bi-level images.

This problem occurs because the binary codes of adjacent integers may differ by several bits. The binary codes of 0 and 1 differ by one bit, those of 1 and 2 differ by two bits, and those of 7 and 8 differ by four bits. The solution is to design special binary codes such that the codes of any consecutive integers i and $i+1$ will differ by one bit only.

Approach 4: Use the context of a pixel to predict its value. The context of a pixel is the values of some of its neighbors. We can examine some neighbors of a pixel P, compute an average A of their values, and predict that P will have the value A. The principle of image compression tells us that our prediction will be correct in most cases, almost correct in many cases, and completely wrong in a few cases. We can say that the predicted value of pixel P represents the redundant information in P. We now calculate the difference

$$\Delta \stackrel{\text{def}}{=} P-A \tag{3-1}$$

and assign variable-length codes to the different values of Δ such that small values (which we expect to be common) are assigned short codes and large values (which are expected to be rare) are assigned long codes. If P can have the values 0 through $m-1$, then values of Δ are in the range $[-(m-1), +(m-1)]$, and the number of codes needed is $2(m-1)+1$ or $2m-1$.

Experiments with a large number of images suggest that the values of Δ tend to be dis-



tributed according to the Laplace distribution[5]. A compression method can, therefore, use this distribution to assign a probability to each value of Δ, and use arithmetic coding to encode the Δ values very efficiently. This is the principle of the MLP method.

The context of a pixel may consist of just one or two of its immediate neighbors. However, better results may be obtained when several neighbor pixels are included in the context. The average A in such a case should be weighted, with near neighbors assigned higher weights. Another important consideration is the decoder. In order for it to decode the image, it should be able to compute the context of every pixel. This means that the context should employ only pixels that have already been encoded. If the image is scanned in raster order, the context should include only pixels located above the current pixel or on the same row and to its left.

Approach 5: Transform the values of the pixels and encode the transformed values. The concept of a transform, as well as the most important transforms used in image compression, is devoted to the wavelet transform. Recall that compression is achieved by reducing or removing redundancy. The redundancy of an image is caused by the correlation between pixels, so transforming the pixels to a representation where they are decorrelated eliminates the redundancy. It is also possible to think of a transform in terms of the entropy of the image. In a highly correlated image, the pixels tend to have equiprobable values, which results in maximum entropy. If the transformed pixels are decorrelated, certain pixel values become common, thereby having large probabilities, while others are rare. This results in small entropy. Quantizing the transformed values can produce efficient lossy image compression. We want the transformed values to be independent because coding independent values makes it simpler to construct a statistical model. We now turn to color images. A pixel in such an image consists of three color components, such as red, green, and blue. Most color images are either continuous-tone or discrete-tone.

Approach 6: The principle of this approach is to separate a continuous-tone color image into three grayscale images and compress each of the three separately, using approaches 3, 4, or 5. For a continuous-tone image, the principle of image compression implies that adjacent pixels have similar, although perhaps not identical, colors. However, similar colors do not mean similar pixel values. Consider, for example, 12-bit pixel values where each color component is expressed in four bits. Thus, the 12 bits 1000 | 0100 | 0000 represent a pixel whose color is a mixture of eight units of red (about 50%, since the maximum is 15 units), four units of green (about 25%), and no blue. Now imagine two adjacent pixels with values 0011 | 0101 | 0011 and 0010 | 0101 | 0011. They have similar colors, since only their red components differ, and only by one unit. However, when considered as 12-bit numbers, the two numbers 001101010011 and 001001010011 are very different,



since they differ in one of their most significant bits.

An important feature of this approach is to use a luminance chrominance color representation instead of the more common RGB[6]. The advantage of the luminance chrominance color representation is that the eye is sensitive to small changes in luminance but not in chrominance. This allows the loss of considerable data in the chrominance components, while making it possible to decode the image without a significant visible loss of quality.

New Words and Phrases

pixel *n.* 像素
encode *vt.* 编码
stream *n.* 流，连续传输的信息序列
zigzag *n.* 锯齿形线条
quad tree 四叉树
redundancy *n.* 冗余，冗余度
nonuniform *adj.* 不均匀的，不一致的
encoder *n.* 编码器
grayscale *n.* 灰度
intuitively *adv.* 直觉地，直观地
adjacent *adj.* 相邻的
decoder *n.* 译码器
correlation *n.* 相关，相关性
eliminate *vt.* 消除，淘汰
equiprobable *adj.* 等概率的
quantize *vt.* 量化，数值化
lossy *adj.* 有损的，致损耗的
luminance *n.* 亮度，发光度
chrominance *n.* 色度

Notes

1. bi-level images：也叫 binary images，二值图像编码。

2. run-length encoding：RLE，游程编码。原理是将一扫描行中的颜色值相同的相邻像素用一个计数值和那些像素的颜色值来代替。

3. space-filling curve：空间填充曲线。

4. Laplace distribution：拉普拉斯分布，如果随机变量的概率密度函数分布为

$$f(x|\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)=\frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & if\ x<\mu\\ \exp\left(-\frac{x-\mu}{b}\right) & if\ x\geqslant\mu\end{cases} \tag{3-2}$$

那么它就是拉普拉斯分布。其中，μ 是位置参数，$b>0$ 是尺度参数。如果 $\mu=0$，那么，正半部分恰好是尺度为 1/2 的指数分布。

5. JBIG：Joint Bi-level Image Experts Group，联合二值图像专家组，是发布二值图像编码标准的专家组。

6. RGB：RGB 色彩模式是工业界的一种颜色标准，是通过对红(R)、绿(G)、蓝(B)

三个颜色通道的变化以及它们相互之间的叠加来得到各式各样的颜色的，RGB 即是代表红、绿、蓝三个通道的颜色，这个标准几乎包括了人类视力所能感知的所有颜色，是目前运用最广的颜色系统之一。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. grayscale image
2. variable-length codes
3. image compression
4. adaptive arithmetic coding
5. wavelet transform
6. pixel values
7. color components
8. redundant information
9. adjacent pixel
10. prefix codes

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

pixel	grayscale	storage	glitches
stream	transmitted	identify	internal
redundant	constructed	corrupted	immune
components	compress	image	misunderstanding
hardware	device	fraction	circuits

1. A possible way to ________ such an image is to scan it, ________ regions, and find repeating regions. If a region B is identical to an already found region A, then B can be compressed by writing a pointer to A on the compressed ________.

2. An image compression method that has been developed specifically for a certain type of image can sometimes be used for other types. Any method for compressing bi-level images, for example, can be used to compress ________ images by separating the bitplanes and compressing each individually, as if it were a bi-level ________.

3. Color images provide another example of using the same compression method across image types. Any compression method for grayscale images can be used to compress

color images. In a color image, each ________ is represented by three color ________(such as RGB).

4. Every time information is ________, over any channel, it may get ________ by noise. In fact, even when information is stored in a ________ device, it may become bad, because no piece of ________ is absolutely reliable.

5. Our language is ________ because only a very small ________ of all possible words are valid. A huge number of words can be ________ with the 26 letters of the English alphabet.

6. Errors are a fact of life. They are all around us, are found everywhere, and are responsible for many ________ and accidents and for much misery and ________. Unfortunately, computer data is not an exception to this rule and is not ________ to errors. Digital data written on a storage ________ such as a disk, CD, or DVD is subject to corruption. Similarly, data stored in the computer's ________ memory can become bad because of a sudden surge in the electrical voltage, a stray cosmic ray hitting the memory ________, or an extreme variation of temperature.

Ⅲ. Translate the following paragraphs into Chinese.

1. Developers and implementers of lossy image compression methods need a standard metric to measure the quality of reconstructed images compared with the original ones. The better a reconstructed image resembles the original one, the bigger should be the value produced by this metric. Such a metric should also produce a dimensionless number, and that number should not be very sensitive to small variations in the reconstructed image.

2. The advantage of the luminance chrominance color representation is that the eye is sensitive to small changes in luminance but not in chrominance. This allows the loss of considerable data in the chrominance components, while making it possible to decode the image without a significant visible loss of quality.

3. A discrete-tone image, also called a graphical image or a synthetic image, is normally an artificial image. It may have a few colors or many colors, but it does not have the noise and blurring of a natural image. Examples are an artificial object or machine, a page of text, a chart, a cartoon, or the contents of a computer screen.

4. A bi-level (or monochromatic) image is an image where the pixels can have one of two values, normally referred to as black and white. Each pixel in such an image is represented by one bit, making this the simplest type of image.

Reading: Pixels

Images are all around us. We see them in color and in high resolution. Many objects (especially artificial objects) seem perfectly smooth, with no jagged edges and no graininess. Computer graphics, on the other hand, deals with images that consist of small dots, pixels. The term pixel stands for "picture element". When we first hear of this feature of computer graphics, we tend to dismiss the entire field as trivial. It seems intuitively obvious that an image that consists of dots would always look pixelated, grainy, rough, and inferior to what we see with our eyes. Yet state-of-the-art computer-generated images are often difficult or impossible to distinguish from their real counterparts, even though they are discrete, made of pixels, and not continuous.

Most engineers, programmers, and users think of pixels as small squares, and this is generally true for pixels on computer monitors. Pixels in other digital output devices (displays or printers) may be rectangular or circular. However, in principle, a pixel should be considered a mathematical, dimensionless, point. It seems impossible to reconstruct a con-



tinuous image from an array of discrete pixels, but this is precisely what the surprising Nyquist-Shannon sampling theorem tells us (in fact, what it guarantees). Here we apply it to two-dimensional images.

Audio is a good starting point to understand the sampling theorem. Sound fed into a microphone is converted to an electrical voltage that varies with time; it becomes a wave. A wave has a frequency, and a wave that varies all the time consists of many frequencies. We denote the maximum frequency contained in a wave by B (cycles per second, or Hertz). The sampling theorem says that it is possible to reconstruct the original wave if it is sampled at a rate greater than $2B$ samples per second.

An image is a rectilinear array of point samples (pixels). The sampling theorem guarantees that we will be able to reconstruct the image (i. e. , to compute the color of every mathematical point in the image) if we sample the image at a rate greater than $2B$ pixels per unit length, where B is the maximum pixel frequency in the image. In practice, pixels, their values, and their frequencies depend on the accuracy of the capturing device. An ideal device should measure the color of an image at certain points, but image sensors (CCDs and CMOS) used in real devices (cameras and scanners) are often far from ideal.

Because of physical limitations, manufacturing defects, and the need to capture enough light, an image sensor often measures the average color (or intensity) of a small area of the image, instead of the color at a point.

Assuming that we have enough pixels for a given digital image, we compute the color of a given point in the image by interpolation. Here, we only touch on the principles of bilinear and bicubic interpolations. The discussion assumes a grayscale image, where each pixel is a number indicating a shade of gray (an intensity), but interpolation can easily be extended to color images, where a pixel is a triplet of primary colors.

Bilinear interpolation. Given enough pixels of an image and given a point Q in the image, we use bilinear interpolation to compute the intensity (grayscale) of the image at Q in the following steps.

1. We select the four pixels surrounding Q, denote their values by a, b, c, and d (Figure 3.1), and convert them to three-dimensional points by considering the value of a pixel the z coordinate of the point and adding appropriate x and y coordinates. We end up with the points $P_{00}=(0, 0, a)$, $P_{01}=(0, 1, b)$, $P_{11}=(1, 1, c)$, and $P_{10}=(1, 0, d)$.

2. We compute the parametric equations of the straight segment $L_1(u)$ running from P_{00} to P_{10} and the segment $L_2(u)$ running from P_{01} to P_{11}. These equations are the simple linear interpolations $L_1(u)=P_{00}(1-u)+P_{10}u$ and $L_2(u)=P_{01}(1-u)+P_{11}u$.

3. We compute the parametric equation Pu, w of the bilinear surface generated by interpolating segments $L_1(u)$ and $L_2(u)$. The equation is



$$P(u, w)=L_1(u)(1-w)+L_2(u)w=P_{00}(1-u)(1-w)+P_{10}u(1-w)+P_{01}(1-u)w+P_{11}uw$$

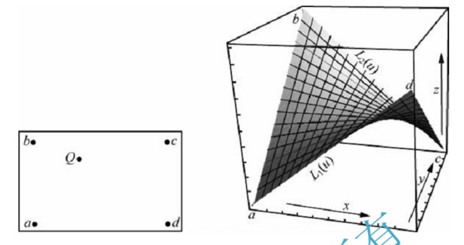


Figure 3. 1　A Bilinear Surface

Figure 3. 1 is an example of such a surface. The entire surface is spanned when parameters u and w are varied independently in the interval [0, 1].

4. We determine the values of u and w for the given point Q and compute its intensity as the z coordinate of surface $P(u, w)$ at these values.

Bicubic interpolation. The principle of bicubic interpolation is similar to that of bilinear interpolation. Given a point Q on the image, we select a group of 4×4 pixels, convert them to three-dimensional points, and compute the bicubic surface $P(u, w)$ that passes through these points. Once this surface is known, the values of parameters u and w for Q are determined and the intensity of image point Q becomes the height of the surface at the point determined by these values.

New Words and Phrases

resolution *n.* 分辨率
artificial *adj.* 人工的，人造的
jagged ages 锯齿边缘
graininess *n.* 粒状
trivial *adj.* 平凡的
intuitively *adv.* 直觉地，直观地
pixelate *vi.* 像素化
　　n. 滤镜
grainy *adj.* 粒状的
state-of-the-art 技术发展水平
counterpart *n.* 副本，配对物
discrete *adj.* 离散的
rectangular *adj.* 矩形的
dimensionless *adj.* 无维数的
voltage *n.* 电压
denote *vt.* 记作，指代，代表
two-dimensional *n.* 二维的，平面的
bicubic *n.* 双三次的，两次立方的
interpolation *n.* 插值
triplet *n.* 三元组
bilinear interpolation 双线性插值(内插)
parametric equation 参数方程



segment *n.* 分割

Exercises

Ⅰ. Answer the following questions.

1. What is pixel?
2. What does sampling theorem guarantee?
3. What is bilinearinterpolation?
4. What is bicubicinterpolation?

Ⅱ. Translate the following sentences into Chinese.

1. Most engineers, programmers, and users think of pixels as small squares, and this is generally true for pixels on computer monitors. Pixels in other digital output devices (displays or printers) may be rectangular or circular. However, in principle, a pixel should be considered a mathematical, dimensionless, point.

2. Because of physical limitations, manufacturing defects, and the need to capture enough light, an image sensor often measures the average color (or intensity) of a small area of the image, instead of the color at a point.

3. Audio is a good starting point to understand the sampling theorem. Sound fed into a microphone is converted to an electrical voltage that varies with time; it becomes a wave. A wave has a frequency, and a wave that varies all the time consists of many frequencies. We denote the maximum frequency contained in a wave by B (cycles per second, or Hertz).



Abstract Reading

Model Correction for Cross-channel Chroma Prediction

A new inter-channel coding mode named LM mode has been intensively explored in the HEVC standardization project. This mode predicts the chroma signal from the luma signal using a linear model whose parameters are inferred from neighboring reconstructed luma and chroma samples. Although this mode presents very good coding efficiency, it is observed that ill cases in the linear parameters calculation can be detected and fixed. This paper gives an overview of the LM mode and presents a novel model correction scheme to detect and correct those ill cases. Simulation results show significant bit-rate savings by the proposed correction scheme with limited added complexity.

Progressive-to-lossless Compression of Color-filter-array Images Using Macropixel spectral-spatial Transformation

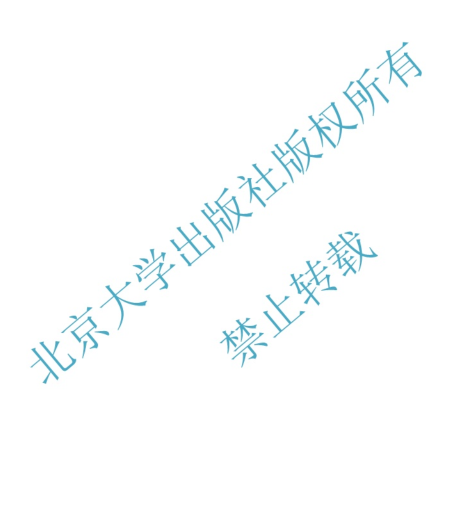
We present a low-complexity integer-reversible spectral-spatial transform that allows for efficient loss less and lossy compression of color-filter-array images (also referred to as camera-raw images). The main advantage of this new transform is that it maps the pixel array values into a format that can be directly compressed in a loss less, lossy, or progressive-to-loss less manner by an existing typical image coder such as JPEG 2000 or JPEG XR. Thus, no special coded design is needed for compressing the camera-raw data. Another advantage is that the new transform allows for mild compression of camera-raw data in a near-loss less format, allowing for very high quality offline post-processing, but with camera-raw files that can be half the size of those of existing camera-raw formats.

Efficient Data Packet Compression for Cache Coherent Multiprocessor Systems

Multiprocessor systems have been popular for their high performance not only for server markets but also for computing environments for general users. With the inreased software complexity, networking overheads in multiprocessor systems are becoming one of the most influential factors in overall system performance. In this paper, we attempt to reduce communication overheads through a data packet compression technique integrating a cache coherence protocol. Here we propose Variable Size Compression (VSC) scheme that



compresses or completely eliminates data packets while harmonizing with existing cache coherence protocols. Simulation results show approximately 23% of improvement on average in terms of overall system performance when compared with the most recent compression scheme. VSC also improves performance by 20% on average in terms of cache miss latency.



UNIT 4

ARTIFICIAL INTELLIGENCE

Text: Properties of Task Environments

The range of task environments that might arise in AIis obviously vast. We can, however, identify a fairly small number of dimensions along which task environments can be categorized. These dimensions determine, to a large extent, the appropriate agent design and the applicability of each of the principal families of techniques for agent[1] implementation. First, we list the dimensions, then we analyze several task environments to illustrate the ideas (Table 4.1). The definitions here are informal.

Table 4.1 Examples of Agent Types and Their PEAS Descriptions

Agent Type	Performance Measure	Environment	Actuators	Sensors
Medical diagnosis system	Healthy patient, reduced costs	Patient, hospital, staff	Display of questions, tests, diagnoses, treatments, referrals	Keyboard entry of symptoms, findings, patient's answers
Satellite image analysis system	Correct image categorization	Downlink from orbiting satellite	Display of scene categorization	Color pixel arrays
Part-picking robot	Percentage of parts in correct bins	Conveyor belt with parts; bins	Jointed arm and hand	Camera, joint angle sensors
Refinery controller	Purity, yield, safety	Refinery, operators	Valves, pumps, heaters, displays	Temperature, pressure, chemical sensors
Interactive English tutor	Student's score on test	Set of students, testing agency	Display of exercises, suggestions, corrections	Keyboard entry

Fully observable vs. partially observable: If an agent's sensors give it access to the complete state of the environment at each point in time, then we say that the task environment



is fully observable. A task environment is effectively fully observable if the sensors detect all aspects that are relevant to the choice of action; relevance, in turn, depends on the performance measure. Fully observable environments are convenient because the agent need not maintain any internal state to keep track of the world. An environment might be partially observable because of noisy and inaccurate sensors or because parts of the state are simply missing from the sensor data—for example, a vacuum agent with only a local dirt sensor cannot tell whether there is dirt in other squares, and an automated taxi cannot see what other drivers are thinking. If the agent has no sensors at all then the environment is unobservable. One might think that in such cases the agent's plight is hopeless, but the agent's goals may still be achievable, sometimes with certainty.

Single agent vs. multiagent: The distinction between single-agent and multiagent environments may seem simple enough. For example, an agent solving a crossword puzzle by itself is clearly in a single-agent environment, whereas an agent playing chess is in a two agent environment. There are, however, some subtle issues. First, we have described how an entity may be viewed as an agent, but we have not explained which entities must be viewed as agents. Does an agent A (the taxi driver for example) have to treat an object B (another vehicle) as an agent, or can it be treated merely as an object behaving according to the laws of physics, analogous to waves at the beach or leaves blowing in the wind? The key distinction is whether B's behavior is best described as maximizing a performance measure whose value depends on agent A's behavior. For example, in chess, the opponent entity B is trying to maximize its performance measure, which, by the rules of chess, minimizes agent A's performance measure. Thus, chess is a competitive multiagent environment. In the taxi-driving environment, on the other hand, avoiding collisions maximizes the performance measure of all agents, so it is a partially cooperative multiagent environment. It is also partially competitive because, for example, only one car can occupy a parking space. The agent-design problems in multiagent environments are often quite different from those in single-agent environments; for example, communication often emerges as a rational behavior in multiagent environments; in some competitive environments, randomized behavior is rational because it avoids the pitfalls of predictability.

Deterministic vs. stochastic: If the next state of the environment is completely determined by the current state and the action executed by the agent, then we say the environment is deterministic; otherwise, it is stochastic. In principle, an agent need not worry about uncertainty in a fully observable, deterministic environment. (In our definition, we ignore uncertainty that arises purely from the actions of other agents in a multiagent environment; thus, a game can be deterministic even though each agent may be unable to predict the actions of the others.) If the environment is partially observable, however, then it

could appear to be stochastic. Most real situations are so complex that it is impossible to keep track of all the unobserved aspects; for practical purposes, they must be treated as stochastic. Taxi driving is clearly stochastic in this sense, because one can never predict the behavior of traffic exactly; moreover, one's tires blow out and one's engine seizes up without warning. The vacuum world as we described it is deterministic, but variations can include stochastic elements such as randomly appearing dirt and an unreliable suction mechanism. We say an environment is uncertain if it is not fully observable or not deterministic. One final note: our use of the word "stochastic" generally implies that uncertainty about outcomes is quantified in terms of probabilities; a nondeterministic environment is one in which actions are characterized by their possible outcomes, but no probabilities are attached to them. Nondeterministic environment descriptions are usually associated with performance measures that require the agent to succeed for all possible outcomes of its actions.

Episodic vs. sequential: In an episodic task environment, the agent's experience is divided into atomic episodes. In each episode the agent receives a percept and then performs a single action. Crucially, the next episode does not depend on the actions taken in previous episodes. Many classification tasks are episodic. For example, an agent that has to spot defective parts on an assembly line bases each decision on the current part, regardless of previous decisions; moreover, the current decision doesn't affect whether the next part is defective. In sequential environments, on the other hand, the current decision could affect all future decisions. Chess and taxi driving are sequential: in both cases, short-term actions can have long-term consequences. Episodic environments are much simpler than sequential environments because the agent does not need to think ahead.

Static vs. dynamic: If the environment can change while an agent is deliberating, then we say the environment is dynamic for that agent; otherwise, it is static. Static environments are easy to deal with because the agent need not keep looking at the world while it is deciding n an action, nor need it worry about the passage of time. Dynamic environments, on the other hand, are continuously asking the agent what it wants to do; if it hasn't decided yet, hat counts as deciding to do nothing. If the environment itself does not change with the passage of time but the agent's performance score does, then we say the environment is semidynamic. Taxi driving is clearly dynamic: the other cars and the taxi itself keep moving while the driving algorithm dithers about what to do next. Chess, when played with a clock, is semidynamic. Crossword puzzles are static.

Discrete vs. continuous: The discrete/continuous distinction applies to the state of the environment, to the way time is handled, and to the percepts and actions of the agent. For example, the chess environment has a finite number of distinct states (excluding the



clock). Chess also has a discrete set of percepts and actions. Taxi driving is a continuous-state and continuous-time problem: the speed and location of the taxi and of the other vehicles sweep through a range of continuous values and do so smoothly over time. Taxi-driving actions are also continuous (steering angles, etc.). Input from digital cameras is discrete, strictly speaking, but is typically treated as representing continuously varying intensities and locations.

Known vs. unknown: Strictly speaking, this distinction refers not to the environment itself but to the agent's (or designer's) state of knowledge about the "laws of physics" of the environment. In a known environment, the outcomes (or outcome probabilities if the environment is stochastic) for all actions are given. Obviously, if the environment is unknown, the agent will have to learn how it works in order to make good decisions. Note that the distinction between known and unknown environments is not the same as the one between fully and partially observable environments. It is quite possible for a known environment to be partially observable—for example, in solitaire[2] card games, I know the rules but am still unable to see the cards that have not yet been turned over. Conversely, an unknown environment can be fully observable—in a new video game, the screen may show the entire game state but I still don't know what the buttons do until I try them.

New Words and Phrases

agent *n.* 智能体
dimension *n.* 维度
detect *vt.* 检查
vacuum *n.* 真空
　　adj. 真空的
multiagent *n.* 多智能体
subtle *adj.* 微妙的，敏感的
entity *n.* 实体，本质
analogous *adj.* 模拟的
opponent *n.* 对手
　　adj. 对立的
collision *n.* 冲突，碰撞
randomize *vt.* 随机化
deterministic *adj.* 确定性的
stochastic *adj.* 随机的
episodic *adj.* 不定期发生的；偶尔发生的
semidynamic *adj.* 半动态的
crossword *n.* 纵横填字字谜
conversely *adv.* 反之，相反地

Notes

1. agent：智能体，指能自主活动的软件或者硬件实体。任何独立的能够思想并可以同环境交互的实体都可以抽象为智能体。



2. solitaire：一个著名的计算机纸牌游戏，需要将两个环状牌基上的纸牌一个按从小到大另一个按从大到小的顺序依次排列。例如，Windows 操作系统中的纸牌大战。

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. discrete distinction
2. continuous distinction
3. sensor
4. single agent
5. multiagent
6. performance measure
7. laws of physics
8. dynamitic environment
9. stochastic environment
10. vacuum agent

Ⅱ. Fill in the blanks with words in the table.

properties	outcome	reduction	environment
location	rationality	distinguish	function
categorization	vacuum	sensors	define
sequence	ensure	perfectu	percept
outcomes	perceiving	preferences	engage

1. Notice that the ________ agent program is very small indeed compared to the corresponding table. The most obvious reduction comes from ignoring the percept history, which cuts down the number of possibilities from 4T to just 4. A further, small ________ comes from the fact that when the current square is dirty, the action does not depend on the ________.

2. We need to be careful to ________ between rationality and omniscience. An omniscient agent knows the actual ________ of its actions and can act accordingly; but omniscience is impossible in reality.

3. How well an agent can behave depends on the nature of the ________; some environments are more difficult than others. We give a crude ________ of environments and



show how ________ of an environment influence the design of suitable agents for that environment.

4. An agent is anything that can be viewed as ________ its environment through ________ and acting upon that environment through actuators. A human agent has eyes, ears, and other organs for sensors and hands, legs, vocal tract, and so on for actuators.

5. By specifying the agent's there is to say about the agent. Mathematically speaking, we say that an agent's behavior is described by the agent ________ that maps any given percept ________ to an action.

6. If we ________ success in terms of agent's opinion of its own performance, an agent could achieve perfect ________ simply by deluding itself that its performance was ________.

7. Our definition of rationality does not require omniscience, then, because the rational choice depends only on the ________ sequence to date. We must also ________ that we haven't inadvertently allowed the agent to ________ in decidedly underintelligent activities.

8. To make such choices, an agent must first have ________ between the different possible ________ of the various plans.

Ⅲ. Translate the following sentences into Chinese.

1. We have listed the medical-diagnosis task as single-agent because the disease process in a patient is not profitably modeled as an agent; but a medical-diagnosis system might also have to deal with recalcitrant patients and skeptical staff, so the environment could have a multiagent aspect. Furthermore, medical diagnosis is episodic if one conceives of the task as selecting a diagnosis given a list of symptoms; the problem is sequential if the task can include proposing a series of tests, evaluating progress over the course of treatment, and so on.

2. Computer scientists are often faced with the task of comparing algorithms to see how fast they run or how much memory they require. There are two approaches to this

task. The first is benchmarking—running the algorithms on a computer and measuring speed in seconds and memory consumption in bytes. Ultimately, this is what really matters, but a benchmark can be unsatisfactory because it is so specific: it measures the performance of a particular program written in a particular language, running on a particular computer, with a particular compiler and particular input data. From the single result that the benchmark provides, it can be difficult to predict how well the algorithm would do on a different compiler, computer, or data set. The second approach relies on a mathematical analysis of algorithms, independently of the particular implementation and input.

____

____

____

____

____

____

____

Reading: Acting Under Uncertainty

Agents may need to handle uncertainty, whether due to partial observability, nondeterminism, or a combination of the two. An agent may never know for certain what state it's in or where it will end up after a sequence of actions.

We have seen problem-solving agents and logical agents designed to handle uncertainty by keeping track of a belief state—a representation of the set of all possible world states that it might be in—and generating a contingency plan that handles every possible eventuality that its sensors may report during execution. Despite its many virtues, however, this approach has significant drawbacks when taken literally as a recipe for creating agent programs.

(1) When interpreting partial sensor information, a logical agent must consider every logically possible explanation for the observations, no matter how unlikely. This leads to impossible large and complex belief-state representations.

(2) A correct contingent plan that handles every eventuality can grow arbitrarily large and must consider arbitrarily unlikely contingencies.

(3) Sometimes there is no plan that is guaranteed to achieve the goal—yet the agent must act. It must have some way to compare the merits of plans that are not guaranteed.



Suppose, for example, that an automated taxi has the goal of delivering a passenger to the airport on time. The agent forms a plan, A_{90}, that involves leaving home 90 minutes before the flight departs and driving at a reasonable speed. Even though the airport is only about 5 miles away, a logical taxi agent will not be able to conclude with certainty that "Plan A_{90} will get us to the airport in time." Instead, it reaches the weaker conclusion "Plan A_{90} will get us to the airport in time, as long as the car doesn't break down or run out of gas, and I don't get into an accident, and there are no accidents on the bridge, and the plane doesn't leave early, and no meteorite hits the car, and..." None of these conditions can be deduced for sure, so the plan's success cannot be inferred. This is the qualification problem, for which we so far have seen no real solution.

Nonetheless, in some sense A_{90} is in fact the right thing to do. What do we mean by this? We mean that out of all the plans that could be executed, A_{90} is expected to maximize the agent's performance measure (where the expectation is relative to the agent's knowledge about the environment). The performance measure includes getting to the airport in time for the flight, avoiding a long, unproductive wait at the airport, and avoiding speeding tickets along the way. The agent's knowledge cannot guarantee any of these outcomes for A_{90}, but it can provide some degree of belief that they will be achieved.

Other plans, such as A_{180}, might increase the agent's belief that it will get to the airport on time, but also increase the likelihood of a long wait. The right thing to do—the rational decision—therefore depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved. The remainder of this section hones these ideas, in preparation for the development of the general theories of uncertain reasoning and rational decisions that we present in this paper.

Summarizing Uncertainty

Let's consider an example of uncertain reasoning: diagnosing a dental patient's toothache. Diagnosis—whether for medicine, automobile repair, or whatever—almost always involves uncertainty. Let us try to write rules for dental diagnosis using propositional logic, so that we can see how the logical approach breaks down. Consider the following simple rule:

$Toothache \Rightarrow Cavity$

The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, an abscess, or one of several other problems:

$Toothache \Rightarrow Cavity \lor Gum\ Problem \lor Abscess \cdots$

Unfortunately, in order to make the rule true, we have to add an almost unlimited list of possible problems. We could try turning the rule into a causal rule:

Cavity ⇒ *Toothache*

But this rule is not right either; not all cavities cause pain. The only way to fix the rule is to make it logically exhaustive: to augment the left-hand side with all the qualifications required for a cavity to cause a toothache. Trying to use logic to cope with a domain like medical diagnosis thus fails for three main reasons.

(1) Laziness: It is LAZINESS too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule and too hard to use such rules.

(2) Theoretical ignorance: Medical science has no complete theory for the domain.

(3) Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

The connection between toothaches and cavities is just not a logical consequence in either direction. This is typical of the medical domain, as well as most other judgmental domains: law, business, design, automobile repair, gardening, dating, and so on. The agent's knowledge an at best provide only a degree of belief in the relevant sentences. Our main tool for PROBABILITY dealing with degrees of belief is probability theory. In the terminology of latter parts, the ontological commitments of logic and probability theory are the same—that the world is composed of facts that do or do not hold in any particular case—but the epistemological commitments are different: a logical agent believes each sentence to be true or false or has no opinion, whereas a probabilistic agent may have a numerical degree of belief between 0 (for sentences that are certainly false) and 1 (certainly true).

Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance, thereby solving the qualification problem. We might not know for sure what afflicts a particular patient, but we believe that there is, say, an 80% chance—that is, a probability of 0.8—that the patient who has a toothache has a cavity. That is, we expect that out of all the situations that are indistinguishable from the current situation as far as our knowledge goes, the patient will have a cavity in 80% of them. This belief could be derived from statistical data—80% of the toothache patients seen so far have had cavities—or from some general dental knowledge, or from a combination of evidence sources.

One confusing point is that at the time of our diagnosis, there is no uncertainty in the actual world: the patient either has a cavity or doesn't. So what does it mean to say the probability of a cavity is 0.8? Shouldn't it be either 0 or 1? The answer is that probability statements are made with respect to a knowledge state, not with respect to the real world. We say "The probability that the patient has a cavity, given that she has a toothache, is 0.8." If we later learn that the patient has a history of gum disease, we can make a different statement: "The probability that the patient has a cavity, given that she has a toothache and a history of gum disease, is 0.4." If we gather further conclusive evidence a-



gainst a cavity, we can say "The probability that the patient has a cavity, given all we now know, is almost 0." Note that these statements do not contradict each other; each is a separate assertion about a different knowledge state.

New Words and Phrases

nondeterminism *n.* 不确定性主义
contingency *n.* 偶发事件；可能性，偶然性
recipe *n.* 方法
hone *n.* 磨石，磨
rational *n.* 有理数
diagnose *vt.* 诊断
diagnosis *n.* 诊断，判断
propositional *adj.* 命题的
cavity *n.* 腔，空洞，洞
gum disease 牙周病
abscess *n.* 脓肿
augment *vt.* 增强，增加
n. 增加，补充物
domain *n.* 论域
antecedent *n.* 前项
adj. 在前的，先行的
exceptionless *adj.* 不除外的，无例外的
commitment *n.* 承诺，许诺
epistemological *adj.* 认识论的，知识论上的
derive *vt.* & *vi.* 得到，导出，来自

Exercises

Ⅰ. Answer the following questions.

1. What is an agent?
2. Why do we need acting under uncertainty?
3. How to summarize uncertainty?

Ⅱ. Translate the following sentences into Chinese.

1. Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance, thereby solving the qualification problem.

____

____

2. Agents may need to handle uncertainty, whether due to partial observability, nondeterminism, or a combination of the two. An agent may never know for certain what state it's in or where it will end up after a sequence of actions.

____

3. When interpreting partial sensor information, a logical agent must consider every logically possible explanation for the observations, no matter how unlikely. This leads to impossible large and complex belief-state representations.

Abstract Reading

An Event-based Distributed Diagnosis Framework Using Structural Model Decomposition

Complex engineering systems require efficient on-line fault diagnosis methodologies to improve safety and reduce maintenance costs. Traditionally, diagnosis approaches are centralized, but these solutions do not scale well. Also, centralized diagnosis solutions are difficult to implement on increasingly prevalent distributed, networked embedded systems. This paper presents a distributed diagnosis framework for physical systems with continuous behavior. Using Possible Conflicts, a structural model decomposition method from the Artificial Intelligence model-based diagnosis (DX) community, we develop a distributed diagnoser design algorithm to build local event-based diagnosers. These diagnosers are constructed based on global diagnosability analysis of the system, enabling them to generate local diagnosis results that are globally correct without the use of a centralized coordinator. We also use Possible Conflicts to design local parameter estimators that are integrated with the local diagnosers to form a comprehensive distributed diagnosis framework. Hence, this is a fully distributed approach to fault detection, isolation, and identification. We evaluate the developed scheme on a four-wheeled rover for different design scenarios to show the advantages of using Possible Conflicts, and generate on-line diagnosis results in simulation to demonstrate the approach.

On the Revision of Informant Credibility Orders

In this paper we propose an approach to multi-source belief revision where the trust or credibility assigned to informant agents can be revised. In our proposal, the credibility of each informant represented as a strict partial order among informant agents, will be maintained in a repository called credibility base. Upon arrival of new information concerning



the credibility of its peers, an agent will be capable of revising this strict partial order, changing the trust assigned to its peers accordingly. Our goal is to formalize a set of change operators over the credibility base: expansion, contraction, prioritized, and non-prioritized revision. These operators will provide the capability of dynamically modifying the credibility of informants considering the reliability of the information. This dynamics will reflect a new perception of trust assigned to the informant, or extend the set of informants by admitting the addition of new informant agents.

How Much Does It Help to Know What She Knows You Know? An Agent-based Simulation Study

In everyday life, people make use of theory of mind by explicitly attributing unobservable mental content such as beliefs, desires, and intentions to others. Humans are known to be able to use this ability recursively. That is, they engage in higher-order theory of mind, and consider what others believe about their own beliefs. In this paper, we use agent-based computational models to investigate the evolution of higher-order theory of mind. We consider higher-order theory of mind across four different competitive games, including repeated single-shot and repeated extensive form games, and determine the advantage of higher-order theory of mind agents over their lower-order theory of mind opponents. Across these four games, we find a common pattern in which first-order and second-order theory of mind agents clearly outperform opponents that are more limited in their ability to make use of theory of mind, while the advantage for deeper recursion to third-order theory of mind is limited in comparison.

UNIT 5

DATA STRUCTURES

Text: Contiguous Representation of Arrays

The obvious way to represent an array in memory is to store its elements in a table, that is, in consecutive cells in memory. For example, consider an array X consisting of six elements X [1] through X [6], where X [i] $= i^2$ for each i. Figure 5.1 shows a contiguous representation of X starting at memory address X, where it is assumed that each integer occupies four memory locations. The ith element of X begins at address $X+4(i-1)$. In general, if X is the address of the first cell in memory of an array with indices $l..u$, and if each element has size L, then the ith element is stored starting at address $X+L(i-1)$ and can be retrieved in constant time.

X	$X+4$	$X+8$	$X+12$	$X+16$	$X+20$
1	4	9	16	25	36
X[1]	X[2]	X[3]	X[4]	X[5]	X[6]

Figure 5.1　A one-dimensional array represented as a table in contiguous storage. The address of the beginning of the array is X; each element occupies four memory locations. The index set of this array is 1,⋯,6 *and* X [i] $=$ i² *for each* i.

What about iteration? It would be possible to iterate over the elements of X by accessing X [l], then X [$l+1$], and so forth up to X [u], thus performing the address calculation Length(X) times. A better method is to start with **X** (which is the address of X [l]) and proceed from element to element by adding L on each iteration. Although this improvement reduces the amount of arithmetic that is performed, the overhead[1] is still linear in the length of the array. Of course, L, l, and u must be available somewhere in order to carry out these calculations. They can be stored in several places.

(1) The values$L$, l, and u can be stored starting at address $\boldsymbol{X}$. The formula for the address of X $[i]$ must then be adjusted slightly to account for the extra space used.

(2) In strongly typed languages, some or all of$L$, l, and u may be part of the definition of X and may be stored elsewhere. Furthermore, if the language does not permit arbitrary lower bounds in indexing then the value I is fixed and need not be stored anywhere.

(3) A sentinel value can be stored just after the last element of the array. That is, memory address $\boldsymbol{X}+L\ (u-l+1)$ can contain some bit pattern that never occurs in the first word of the memory representation of any element of V. Now u need not be explicitly stored at all and iterations are terminated by detecting the sentinel value. A disadvantage of this method is that an iteration is required even to find the length of such an array. Nevertheless, this representation is often used when L and l are fixed. The programming language C, for example, represents character strings in this way.

Storage in contiguous memory is less attractive when the elements of the array have different lengths, because the ith element cannot be found in constant time by simple arithmetic. To handle this situation we can store the elements in memory anywhere and keep a table of pointers to the elements. Figure 5. 2 shows an example of such an array whose elements in order are the integer 9, the string $ABCD$, and an array of two integers. (The latter two arrays are stored in contiguous memory, not as tables of pointers.) The address of the ith element is now stored in location $X+P$. $(i-1)$ where P is the size of a pointer in memory. The disadvantages of this implementation are two: an extra pointer must be followed to perform an Access, and an array of length n uses $p\times n$ extra bits to store pointers in addition to the space needed to store the data. But a major advantage is that single pointer manipulations suffice to move elements within the array; for this reason, tables of pointers are often used when the elements are large (even if they are all of the same size).

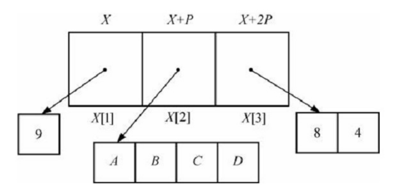


Figure 5. 2 A three-element array implemented as a table of pointers.

A two-dimensional array whose elements all have the same size can also be represented efficiently in contiguous storage; the only problem is to determine the order in which the elements should be placed. The two most common schemes are row major order, in which

the rows are placed one after another in memory, and column major order, in which the columns are placed one after another. For example, consider the following two arrays, each of which has indices $(1\cdots4)\times(1\cdots5)$:

$$R=\begin{pmatrix}1&2&3&4&5\\6&7&8&9&10\\11&12&13&14&15\\16&17&18&19&20\end{pmatrix}\qquad C=\begin{pmatrix}1&5&9&13&17\\2&6&10&14&18\\3&7&11&15&19\\4&8&12&16&20\end{pmatrix}\tag{5-1}$$

The entries in array R suggest the order in which the elements of R are stored in row major order. First comes $R[1,1]$, then $R[1,2]$, and so forth up to $R[1,5]$ which is followed by $R[2,1]$. In general, entry $R[i,j]$ is stored in memory at address $\mathbf{R}+L\cdot(5(i-1)+(j-1))$, where as usual each element requires space L and the first element begins at address $\mathbf{R}$. (The subtractions here reflect the fact that 1 is the first integer in each interval indexing R.) The entries in C suggest column major order. Again element $C[1,1]$ is first in memory, but it is followed by $C[2,1]$, $C[3,1]$, $C[4,1]$ and then $C[1,2]$. If C is stored in column major order, entry $C[i,j]$ begins at address $\mathbf{C}+L\cdot(4(j-1)+(i-1))$. If particular iterations are anticipated—for example, if row-by-row iteration is more frequent than column-by-column iteration—then one of these layouts may be more advantageous than the other.

Row and column major[2] order can be generalized to higher dimensions. Let X be a general d-dimensional array with indices $(l_1\cdots u_1)\times\cdots\times(l_d\cdots u_d)$. When X is stored in row major order the first element is $X[l_1,\cdots,l_d]$, followed by $X[l_1,\cdots,l_{d+1}]$, $X[l_1,\cdots,l_{d+2}]$, and so forth up to $X[l_1,\cdots,l_{d-1},u_d]$, after which the next element is $X[l_1,\cdots,l_{d-1}+1,l_d]$. When arrays are represented in row major order we often say simply that "the last index varies fastest" as we examine successive elements in memory; each index is incremented only after all subsequent indices reach their upper bounds. Similarly, to store X in column major order we store $X[l_1,l_2,\cdots,l_d]$, $X[l_1+1,l_2,\cdots,l_d]$, and so forth up to $X[u_1,l_2,\cdots,l_d]$, and the next element is $X[l_1,l_2+1,l_3,\cdots,l_d]$, so that it is the first index that "varies fastest".

Now suppose X is an arbitrary d-dimensional array as in the previous paragraph, that X is stored in row major order starting at address $\mathbf{X}$, and that each element of X occupies space L. For arbitrary indices $j_1,j_2,\cdots,j_d$, where in memory is element $X[j_1,j_2,\cdots,j_d]$ located? (Of course, the answer is "nowhere" unless $l_k\leqslant j_k\leqslant u_k$ for each k. Verifying this condition is called range checking. Not all languages perform range checking; in some, it can be turned on for debugging and turned off when efficiency is important.)

For each $k=0,1,\cdots,d$, define $M_k=L_{s_{k+1}\cdots s_d}$ where $s_i=u_i-l_i+1$ is the size of the ith dimension of X. M_k is the amount of memory required to store each $d-k$dimensional "subarray" of X in which the first k indices are fixed; for example, M_k is the number of memory



locations from the start of element $X[l_1, l_2, \cdots, l_d]$ to the end of element $X[l_1, l_2, \cdots, l_k, u_{k+1}, u_{k+2}, \cdots, u_d]$. In particular, $M_d = L$ and M_0 is the size of the entire array X. Therefore, there are $M_1 \cdot (j_1 - l_1)$ cells from X to the beginning of element $X[j_1, l_2, \cdots, l_d]$. From that point, there are $M_2 \cdot (j_2 - l_2)$ cells to the beginning of element $X[j_1, j_2, l_3, \cdots, l_d]$, Continuing in this way, we find that element $X[j_1, j_2, \cdots, j_d]$ is located at address

$$X + M_1(j_1 - l_1) + M_2(j_2 - l_2) + \cdots + M_d(j_d - l_d) \tag{5-2}$$

To make the Access operation as fast as possible, the values M_k should be computed in advance, once and for all. Moreover, we should compute and save the single constant value $X_0 = M_1 l_1 + \cdots + M_d l_d$, since then we can write expression (1) as

$$X - X_0 + M_1 j_1 + M_2 j_2 + \cdots + M_d j_d \tag{5-3}$$

which is faster to evaluate, requiring only approximately $2d$ operations rather than $3d$ operations. Note that once we have the M_k and X_0, the l_k and u_k are unnecessary for Access unless we wish to perform range checking.

There is an independent context in which the M_k can be useful: as mentioned before, row major representation of X is especially appropriate when we desire to iterate over the elements of X with the last index varying fastest. Suppose that we wish to have a version of the Iterate operation with the indices changing in some other order. Any such iteration can be implemented efficiently using the fact that the distance in memory between $X[j_1, \cdots, j_k, \cdots, j_d]$ and $X[j_1, \cdots, j_k+1, \cdots, j_d]$ is exactly M_k.

When the elements of a multidimensional array are of different sizes in memory, we can extend the scheme of Figure 5.2 by storing pointers to the elements rather than the elements themselves. Then L is equal to the size of a pointer, and a pointer must be followed after the address calculation.

All of the methods so far considered for representing multidimensional arrays permit access to any element of the array in constant time. There is a subtle point here. You may feel that access to an element of a multidimensional array cannot be performed in constant time, since the number of arithmetic operations depends on the number of dimensions. But the "size" of an array is the total number of its elements; the cost of accessing any element of a d-dimensional array is independent of the number of elements in the array, although it does depend on d. This convention reflects the fact that the arrays used in computer programs typically have a fixed number of dimensions, although they may have more or fewer elements depending on the problem size. Indeed, few languages support arrays in which d is not fixed for each given array.

New Words and Phrases

consecutive *adj*. 连续的，连贯的
contiguous *adj*. 紧密连接的
indices *n*. index的复数；索引
retrieve *vt*. 检索
linear *adj*. 线性的，一次的
sentinel *n*. 哨兵
terminate *vt*. & *vi*. 结束，使终结
manipulations *n*. 操作，操纵
suffice *vi*. 足够；有能力
vt. 满足……的需要；使满足
iteration *n*. 迭代
debug *vt*. 排除故障，程序调试命令
subarray *n*. 子数组
iterate *vt*. 重复；反复申明
convention *n*. 约定，惯例
increment *n*. 增量，增长

Notes

1. overhead：额外开销。除了有用数据以外，还有很多其他信息，这些信息用来保证数据结构。这些信息被称作额外开销。

2. row and column major order：行主序和列主序。

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. one-dimensional array
2. sentinel value
3. three-element array
4. contiguous memory
5. row-by-row iteration
6. column-by-column iteration
7. multidimensional array
8. the programming language C



Ⅱ. Fill in the blanks with the missing word(s) from the table below.

applications	representation	scattering	median
successively	additional	contiguous	value
element	string	remaining	binary
consists	drawbacks	linear	subtree
volumes	deleted	compressed	associated

1. One of the ________ of representing arrays in ________ memory is the time required to initialize them; the obvious method of ________ setting each element to its initial value uses time proportional to the number of elements.

2. The contiguous ________ methods of the preceding section allocate storage for every ________ of an array. But in many ________ the arrays under consideration are only partially filled. Sometimes only a ________ of the elements of an array have useful values.

3. Decoding is almost the same as encoding. First of all, the ________ representation ________ simply of a sequence of code numbers; it is easy to retrieve them one by one since the length in bits of a single code number is fixed.

4. Retrieval of information from large text files is a very broad and important problem-numerous techniques have been developed and entire ________ written on the topic. One simple aspect of this problem is ________ searching.

5. A binary search tree is a binary tree having a ________ associated with each node, such that the values have a ________ order, and at each node the value is greater than any value in the left subtree and less than any value in the right ________.

6. When a node is ________ from a binary search tree, the inorder traversal of the ________ nodes must yield the keys in the same order they had before the deletion.

7. In a ________ split tree each node contains two keys. One is called the node value; any ________ information in the node (the Info field) is associated with this key.

8. The data structure is a pure search tree—the nodes have no balance, color, or other auxiliary fields, only left and right child pointers and fields for the key itself and any ________ data.

Ⅲ. Translate the following paragraphs into Chinese.

1. A greedy algorithm is based on the following simple principle: when called on to make a sequence of choices to develop a solution to a problem, always make the choice that has the lowest immediately visible cost; don't try to look ahead to see if that choice might

turn out to be more costly in the long run. The first thing to note about greedy algorithms is that, for most problems (as in most real-life situations), they do not work.

2. Algorithms that solve very different problems sometimes bear a strong family resemblance to each other. For example, the general strategy of Divide-and-Conquer underlies Binary Search, Merge Sort, and the clever integer multiplication algorithm, as well as useful algorithms for many other problems. It is worth keeping the general idea of Divide-and-Conquer in mind when faced with new problems to solve. In this section we will give examples of several other general strategies for the design of algorithms.

Reading: Programming as an Engineering Activity

A program is a solution to a problem. The problem might be very specific and well-defined, for example, to calculate the square roots of the integers from 1 to 100 to ten decimal places. Or the problem might be vast and vague, for example, to develop a system for printing books by computer. Large, ill defined problems are, however, best solved by breaking them down into smaller and more specific problems. As a part of the problem of printing books by computer, for example, we might need to determine the places where a word could be hyphenated if it had to be split across two lines. Our subject matter is programming problems that are specific enough that we can describe them in a few words and can judge readily what is a solution and what isn't, but are common enough that they come up over and over again in the solution of larger programming problems.

Even for problems that can be described very exactly in a few words, however, there can be many possible solutions. Of course one can always get different programs by changing variable names, translating from FORTRAN to Pascal, and the like. But there can be solutions that differ in more fundamental ways, that use quite different approaches or



methods to solve a problem. Consider, for example, the problem of finding a word K in a sorted table of words.

Here are three approaches.

A. Start at the beginning of the table and go through it, comparing K to each word in the table, until you find K or reach the end of the table. Of course that way doesn't take advantage of the fact that the table is sorted.

Here's a slightly more intelligent variation.

B. Start at the beginning of the table and go through it as in (A), stopping when you find K or another word that should come after K in the table, or when you reach the end of the table.

Changing the stopping condition in this way eliminates some unnecessary work done by method (A). If we're looking for **a ardvark**, for example, chances are we won't have to look long if we use method (B). But there is a better way yet.

C. Start in the middle of the table. If K is the middle word in the table, you're done. Otherwise, decide by looking at that middle word whether K would be in the first half of the table or the second, and repeat the same process on one half of the table. On subsequent iterations search a quarter, an eighth, …of the table in the same way. Stop when you find K or have shrunk to nothing the size of the table you're searching.

Method (C) is called binary search and is generally the fastest of the three. (It's also the trickiest to program correctly. Actually, this description leaves out a lot of important details; for example, which element is in the "middle" of a table of length 10?) We'll get to a detailed account of binary search in the previous papers, but for now there are a few morals to be drawn from the example. First, (A), (B), and (C) are different algorithms for the same problem. None of them is a program, since the language used to describe them isn't a programming language. But any programmer would understand these descriptions, and would understand that FORTRAN and Pascal implementations of (C) embody the same algorithm, whereas Pascal implementations of (A) and (C) embody utterly different algorithms.

An algorithm is a computational method used for solving a problem. The goals of this book are to teach you some of the most important algorithms for solving problems that come up over and over again in computer programming, and to teach you how to decide which algorithm to use when you have a choice (as you almost always do).

We might choose one algorithm over another because it is always faster, or because it is usually faster, or because it uses less memory. Or we might choose an algorithm because it is easier to program, or because it is more general and we want to anticipate the possibili-

ty that the problem we are solving might change in the future. For our purposes in this book, however, we will mostly be looking at the speed of algorithms, and how much memory they use.

Of course we are not going to determine the speed of an algorithm by writing a program and then timing it. The numbers obtained in this way would depend too much on the quality of the programmer and the speed of the particular computer to be of general interest or applicability. Instead, we'll try to think in more abstract, mathematical terms. If the table has length n, then method (A) takes time proportional to n; double the size of the table and the algorithm will take roughly twice as long. Method (C), on the other hand, takes time proportional to the base 2 logarithm of n at worst (since that is the number of times you can divide a table of length n in half before it is reduced to a single element).

We'll spend a good deal of time on this business of algorithm analysis, but again a few simple morals will suffice for now. We want to use mathematical tools for analyzing the algorithms we consider, since the right mathematical tools will give us conclusions that hold for all implementations.

To develop those mathematical tools, we have to come up with mathematical models for the situations we are trying to understand. For example, to conclude that method (A) takes time proportional to the length of the table, we need assume only that it always takes the same amount of time to get from any element of the table to the next. That is true for a great many ways of implementing tables, so from a weak assumption we can draw a conclusion of quite general applicability.

Programming is an engineering activity. It isn't pure science, or pure mathematics either; when we write programs, we can't ignore annoying details of practical importance, and we're not working in an environment where there's only one right answer. Engineers make design decisions based on an understanding of the consequences of alternative choices. That understanding comes from a knowledge of laws, usually stated in mathematical terms, that cover a broad variety of situations. An engineer decides what kind of bridge to build to span a river at a particular spot by sizing up the parameters of the situation (how long? how much weight to be borne?) and applying the general laws that characterize the behavior of various kinds of bridges. An engineer will also bring to bear the wisdom of experience accumulated by witnessing the construction of the things that have been designed. Programmers should think the same way; they need both an understanding of the general laws that govern the performance of algorithms, and the practical wisdom that comes from having attempted to implement them.



New Words and Phrases

square roots *n*. 平方根
vague *adj*. 模糊的
hyphenate *vt*. 以连字符号连接
aardvark *n*. 非洲食蚁兽
proportional *adj*. 成比例的
witness *vi*. 目击，见证

Exercises

Ⅰ. Answer the following questions.

1. How to find a word *K* in a sorted table of words?
2. Why programming is an engineering activity?
3. Why algorithm is important to programming?

Ⅱ. Translate the following sentences into Chinese.

1. A program is a solution to a problem. The problem might be very specific and well-defined-for example, to calculate the square roots of the integers from 1 to 100 to ten decimal places. Or the problem might be vast and vague, for example, to develop a system for printing books by computer.

____

____

____

____

2. As a part of the problem of printing books by computer, for example, we might need to determine the places where a word could be hyphenated if it had to be split across two lines. Our subject matter is programming problems that are specific enough that we can describe them in a few words and can judge readily what is a solution and what isn't, but are common enough that they come up over and over again in the solution of larger programming problems.

____

____

____

____

____

3. Programming is an engineering activity. It isn't pure science, or pure mathematics either; when we write programs, we can't ignore annoying details of practical importance, and we're not working in an environment where there's only one right answer.

Abstract Reading

Finding Dominators via Disjoint Set Union

The problem of finding dominators in a directed graph has many important applications, notably in global optimization of computer code. Although linear and near-linear-time algorithms exist, they use sophisticated data structures. We develop an algorithm for finding dominators that uses only a "static tree" disjoint set data structure in addition to simple lists and maps. The algorithm runs in near-linear or linear time, depending on the implementation of the disjoint set data structure. We give several versions of the algorithm, including one that computes loop nesting information (needed in many kinds of global code optimization) and that can be made self-certifying, so that the correctness of the computed dominators is very easy to verify.

Strict Fibonacci Heaps

We present the first pointer-based heap implementation with time bounds matching those of Fibonacci heaps in the worst case. We support make-heap, insert, find-min, meld and decrease-key in worst-case O(1) time, and delete and delete-min in worst-case O(lg n) time, where n is the size of the heap. The data structure uses linear space.

A previous, very complicated, solution achieving the same time bounds in the RAM model made essential use of arrays and extensive use of redundant counter schemes to maintain balance. Our solution uses neither. Our key simplification is to discard the structure of the smaller heap when doing a meld. We use the pigeonhole principle in place of the redundant counter mechanism.

Deletion without Rebalancing in Balanced Binary Trees

We address the vexing issue of deletions in balanced trees. Rebalancing after a deletion



is generally more complicated than rebalancing after an insertion. Textbooks neglect deletion rebalancing, and many database systems do not do it. We describe a relaxation of AVL trees in which rebalancing is done after insertions but not after deletions, yet access time remains logarithmic in the number of insertions. For many applications of balanced trees, our structure offers performance competitive with that of classical balanced trees. With the addition of periodic rebuilding, the performance of our structure is theoretically superior to that of many if not all classic balanced tree structures. Our structure needs O(log log m) bits of balance information per node, where m is the number of insertions, or O(log log n) with periodic rebuilding, where n is the number of nodes. An insertion takes up to two rotations and O(1) amortized time. Using an analysis that relies on an exponential potential function, we show that rebalancing steps occur with a frequency that is exponentially small in the height of the affected node.

UNIT 6

INFORMATION SECURITY

Text: Psychological Security Traps

During my career of attacking software and the facilities they power, many colleagues have remarked that I have a somewhat nonstandard approach. I tended to be surprised to hear this, as the approach seemed logical and straightforward to me. In contrast, I felt that academic approaches were too abstract to realize wide success in real-world applications. These more conventional disciplines were taking an almost completely random tack with no focus or, on the opposite end of the spectrum, spending hundreds of hours reverse-engineering and tracing applications to (hopefully) discover their vulnerabilities before they were exploited out in the field.

Now, please do not take this the wrong way. I'm not condemning the aforementioned techniques. In fact I agree they are critical tools in the art of vulnerability discovery and exploitation. However, I believe in applying some shortcuts and alternative views to envelope, enhance, and—sometimes—by pass these approaches.

In this article I'll talk about some of these alternative views and how they can help us get inside the mind of the developer whose code or system we engage as security professionals. Why might you want to get inside the mind of the developer? There are many reasons, but for this article we will focus on various constraints that are imposed on the creation of code and the people who write it. These issues often result in suboptimal systems from the security viewpoint, and by understanding some of the environmental, psychological, and philosophical frameworks in which the coding is done, we can shine a spotlight on which areas of a systemare more likely to contain vulnerabilities that attackers can exploit. Where appropriate, I'll share anecdotes to provide examples of the mindset issue at hand.

My focus for the past several years has been on large-scale environments such as major



corporations, government agencies and their various enclaves, and even nation states. While many of the elements are applicable to smaller environments, and even to individuals, I like to show the issues in larger terms to offer a broader social picture. Of course, painting with such a broad brush requires generalizations, and you may be able to find instances that contradict the examples.

The goal here is not to highlight particular technologies, but rather to talk about some environmental and psychological situations that caused weak security to come into being. It is important to consider the external influences and restrictions placed on the implementers of a technology, in order to best understand where weaknesses will logically be introduced. While this is an enjoyable mental game to play on the offensive side of the coin, it takes on new dimensions when the defenders also play the game and a) prevent errors that would otherwise lead to attacks or b) use these same techniques to game the attackers and how they operate[1].

At this point, the security game becomes what I consider beautiful. The mindsets I'll cover fall into the categories of learned helplessness and naïveté, confirmation traps, and functional fixation. This is not an exhaustive list of influencing factors in security design and implementation, but a starting point to encourage further awareness of the potential security dangers in systems that you create or depend on.

Learned Helplessness and Naïveté

Sociologists and psychologists have discovered a phenomenon in both humans and other animals that they call learned helplessness. It springs from repeated frustration when trying to achieve one's goals or rescue oneself from a bad situation. Ultimately, the animal subjected to this extremely destructive treatment stops trying. Even when chances to do well or escape come along, the animal remains passive and fails to take advantage of them.

To illustrate that even sophisticated and rational software engineers are subject to this debilitating flaw, I'll use an example where poor security can be traced back to the roots of backward compatibility. Backward compatibility is a perennial problem for existing technology deployments. New technologies are discovered and need to be deployed that are incompatible with, or at the very least substantially different from, existing solutions.

At each point in a system's evolution, vendors need to determine whether they will forcibly end-of-life the existing solutions, provide a migration path, or devise a way to allow both the legacy and modern solutions to interact in perpetuity. All of these decisions have numerous ramifications from both business and technology perspectives. But the decision is usually driven by business desires and comes down as a decree to the developers and engineers. When this happens, the people responsible for creating the

actual implementation will have the impression that the decision has already been made and that they just have to live with it. No further reevaluation or double guessing need take place.

Imagine that the decision was made to maintaincompatibility with the legacy technology in its replacement. Management further decrees that no further development or support work will take place on the legacy solution, in order to encourage existing customers to migrate to the replacement.

Although such decisions place burdens on the development in many ways—with security implications—they are particularly interesting when one solution, usually the new technology, is more secure than the other. In fact, new technologies are often developed explicitly to meet the need for greater security—and yet the old technology must still be supported. What security problems arise in such situations?

There are different ways to achieve backward compatibility, some more secure than others. But once the developers understand that the older, less secure technology is allowed to live on, solutions that would ease the risk are often not considered at all. The focus is placed on the new technology, and the legacy technology is glued into it (or vice versa) with minimal attention to the legacy's effects. After all, the team that is implementing the new technology usually didn't develop the legacy code and the goal is to ultimately supplant the legacy solution anyway—right?

The most direct solution is to compromise the robustness and security strength of the new technology to match that of the legacy solution, in essence allowing both the modern and legacy technology to be active simultaneously. Learned helplessness enters when developers can't imagine that anything could be done—or worse, even should be done—to mitigate the vulnerabilities of the legacy code. The legacy code was forced on them, it is not perceived to be their bailiwick (even if it impacts the security of the new technology by reducing it to the level of the old), and they feel they are powerless to do anything about it anyway due to corporate decree.

New Words and Phrases

spectrum *n.* 谱
aforementioned *adj.* 上述
suboptimal *adj.* 次优的
spotlight *n.* 聚焦
system *n.* 系统
anecdote *n.* 轶事
enclave *n.* 其他领域；飞地(指在本国境内的隶属另一国的一块领土)
contradict *vt.* 与……矛盾
implementer *n.* 实施者
naïveté *adj.* 天真的
springs *vi.* 使……裂开



perennial *adj*. 多年的，长久的
incompatible *adj*. 不匹配的，不相容的
end-of-life *n*. 报废
perpetuity *n*. 永久
ramification *n*. 结果，分支
compatibility *n*. 兼容性
legacy technology 传统技术
implication *n*. 影响
explicitly *adv*. 明确地
minimal *adj*. 最小的，极少的；极小的
supplant *vt*. 把……排挤掉，取代；代替
robustness *n*. 坚固性，健壮性；鲁棒性
bailiwick *n*. 工作区间；辖区
decree *n*. 法令，命令；(法院的)判决，裁定

Notes

1. While this is an enjoyable mental game to play on the offensive side of the coin, it takes on new dimensions when the defenders also play the game and a) prevent errors that would otherwise lead to attacks or b) use these same techniques to game the attackers and how they operate.

作为进攻方玩这个心理游戏是很有意思的。当防守方也玩这个游戏并且避免会招致攻击的错误或者运用这些相同的技巧来戏弄攻击者及他们的操作时，这个游戏就会上升到一个新的维度。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. random tack
2. reverse-engineering
3. tracing application
4. functional fixation
5. confirmation trap
6. debilitating flaw
7. migration path
8. legacy technology
9. backward compatibility

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

hash	compatibility	valid	capability
negative	confirmation	agency	multiple
alternative	embodied	refute	password
proximity	notion	ensuring	fixation
corporate	inability	solution	permanently

1. In short, the problems of the new modern security ________ sprang from the weaker LANMAN ________ of the legacy system and thus reduced the entire security profile to its lowest common denominator. It wasn't until much later, and after much ________ security publicity, that Microsoft introduced the ________ of sending only one hash or the other, and not both by default—and even later that they stopped storing both LANMAN and NT hashes in ________ to each other on local systems.

2. The issue from a security standpoint becomes how to accomplish this backward ________ without degrading the security of the new systems. Microsoft's naïve solution ________ pretty much the worst of all possibilities: it stored the insecure ________ together with the more secure one, and for the benefit of the attacker it transmitted the representations of both hashes over the network, even when not needed.

3. Consider an intelligence analyst working at a three-letter ________. The analyst wants to create useful reports in order to progress up the career ladder. The analyst culls information from ________ sources, including the previous reports of analysts in her position.

4. Functional fixation is the ________ to see uses for something beyond the use commonly presented for it. This is similar to the ________ of first impressions—that the first spin applied to initial information disclosure (e. g. , a biased title in a newspaper report or a presentation of a case by a prosecutor) often ________ influences the listener's ongoing perception of the information.

5. One of the greatest hampers to security springs from negative perceptions of security requirements at a high ________ level. Some of these represent functional ________.

6. We can overcome learned helplessness and naïveté by ________ that initial decisions do not shut off creative thinking.

7. We can overcome ________ traps by seeking inputs from diverse populations and forcing ourselves to try to ________ assumptions.

8. We can overcome functional fixation by looking for ________ uses for our tools, as



well as alternative paths to achieve our goals.

Ⅲ. Translate the following paragraphs into Chinese.

1. Many people think of security products such as vulnerability scanners and anti-virus software as tools that increase the security of a system or organization. But if this is the only view you hold, you are suffering from functional fixation. Each of these technologies can be very complex and consist of thousands of lines of code. Introducing them into an environment also introduces a strong possibility of new vulnerabilities and attack surfaces.

____

____

____

____

____

2. Modern anti-virus software, unfortunately, has been found to include all sorts of common programming vulnerabilities, such as local buffer overflows, unchecked execution capabilities, and lack of authentication in auto-update activities. This security software, therefore, can also become the opening for attackers rather than the defense it was intended for.

____

____

____

____

____

3. Microsoft preferred for all their customers to upgrade to newer versions of Windows, of course, but did not dare to cut off customers using older versions or even retrofit them with the new hash function. Because the password was a key part of networking, they had to assume that, for the foreseeable future, old systems with no understanding of the new hash function would continue to connect to systems fitted out with the more secure hash.

____

____

____

____

____



Reading: Sunk Costs versus Future Profits: An Energy Example

Part of my career has involved examining in detail the back end control systems at various electric utilities and, to a somewhat lesser extent, oil company backend systems. I assessed how they were protected and traced their interconnections to other systems and networks. It was surprising how the oil and electric industries, while using such similar systems and protocols, could be operated and run in such widely disparate configurations and security postures.

To put it politely, the electric company networks were a mess. Plant control systems and networks could be reached from the public Internet. General-purpose systems were being shared by multiple tasks, interleaving word processing and other routine work with critical functions that should have been relegated to specialized systems to prevent potential interference or disruption of operations. It appeared in several cases that systems and networks had been put together on a whim and without consideration of optimal or even accurate operations. Implementers moved on to the next job as soon as things worked at all. Many plant control networks, plant information networks, and corporate LANs had no firewalls or chokepoints. From a security standpoint, all this combined to create the potential for malicious interlopers to wreak serious havoc, including manipulating or disrupting the physical components used in the production and transmission of power.

Conversely, the few offshore oil systems that I had looked at, while utilizing similar SCADA systems, were configured and operated in a different fashion. Plant control and information networks were strictly segregated from the corporate LAN. Most critical systems were set correctly to have their results and status handled by a librarian system that then pushed the information out in a diode fashion to higher analysis systems. Concise and efficient network diagrams resulted in crisp and clean implementations of SCADA and DCS systems in the physical world, including restriction of access that resulted in effective security. In many cases the components were custom systems designed and configured to perform only specific functions.

The contrast between the electric and oil organizations intrigued and worried me. As fate would have it, I was in the position to be able to call a meeting about this subject with some high ranking technical people from electric companies, oil companies, and government (think spook) agencies.

The first salient aspect that surprised me from the meeting was that the people from the electric utilities and their electric utility oversight and clearinghouse organizations did



not refute my statements regarding the poor—or completely missing—security on their networks and systems. This surprised me because the electric companies were publicly denying that they had any cyber-system risk. In our meeting they pointed out some examples where security had been implemented correctly—but they acknowledged that these examples were exceptions and not the norm.

My second surprise came when the oil companies stated that they did not go about designing their systems from a security perspective at all, and that although security was important, it was not the business driver for how things were configured. The primary driver was to have an edge against their direct competitors. If company A could make a critical component operate at 5% greater efficiency than company B, the increased operational capacity or reduction in overhead rewarded company A over time with large sums of money. Examples of how to increase such efficiency included:

(1) Forced separation and segregation of systems to prevent critical systems from incurring added latency from being queried by management and reporting requests.

(2) Utilizing special-purpose systems designed to accomplish specific tasks in place of general purpose nonoptimized systems.

These efficiencies benefited security as well. The first created strong, clean, and enforceable boundaries in networks and systems. The second produced systems with smaller surface areas to attack.

Enforceable network and system boundaries are an obvious effect, but the case of smaller surface areas deserves a brief examination. Imagine that you have a general-purpose system in its default configuration. The default configuration might have several services already configured and running, as well as many local daemons executing to assist user processing. This allows the system to be deployed in the largest number of settings with minimal reconfiguration required. The systems' vendor prefers such systems with broad capabilities because they make installation easier.

However, this doesn't mean that the default setting is optimal for the majority of consumers, just that it is acceptable. In the default setting, each of the running services is an attack surface that may be exploited. Similarly, client applications may be compromised through malicious input from compromised or falsified servers. The more services and client applications that are running on the system, the greater the attack surface and the greater the likelihood that the system can be remotely or locally compromised.

Having a large attack surface is not a good thing, but the drawback of generality examined by the oil companies was the systems' suboptimal performance. For each running program, which includes server services as well as local applications, the kernel and CPU devotes processing time. If there are many running applications, the system has to time-

slice among them, a kernel activity that in itself eats up resources.

However, if there are few running applications, each one can have a greater number of CPU slices and achieve greater performance. A simple way to slim down the system is to remove superfluous services and applications and optimize the systems to run in the most stripped-down and dedicated fashion possible. Another way is to deploy systems dedicated to specific functions without even the capability of running unrelated routines. These tactics had been used by the oil companies in the offshore rigs I had examined in order to maximize performance and thus profits.

Why hadn't the electric utilities gone through the same exercise as the oil companies? At first, electric companies were regulated monopolies. Where these companies did not need to be competitive, they had no drive to design optimized and strictly structured environments.

One would be tempted to assume that deregulation and exposure of electric companies to a competitive environment would improve their efficiency and (following the same path as oil companies) their security. However, the opposite occurred. When the electric companies were turned loose, so to speak, and realized they needed cost-cutting measures to be competitive, their first steps were to reduce workforce. They ended up assigning fewer people to maintain and work on the same number of local and remote systems (often through remote access technologies), focusing on day-to-day operations rather than looking ahead to long-term needs. This is usually a poor recipe for efficiency or security.

The story of the oil companies confirms the observation I made in the previous section about the ISP. Most organizations think of security as a sunk cost, insofar as they think of it at all. Security approached in this fashion will likely be inadequate or worse. If, however, one focuses on optimizing and streamlining the functionality of the networks and systems for specific business purposes, security can often be realized as a by-product. And once again, security professionals can further their cause by overcoming their functional fixation on security as a noble goal unto itself worth spending large sums on, and instead sometimes looking at sneaking security in as a fortuitous by-product.

New words and Phrases

backend *n.* 后端

protocol *n.* (数据传递的)协议

configuration *n.* (计算机的)配置

disruption *n.* 分裂，瓦解；破裂，毁坏；中断

implementer *n.* 实施者

chokepoint *n.* 阻塞点

malicious *adj.* 恶意的，有敌意的；蓄意的；预谋的；存心不良的

havoc *n.* 大破坏，浩劫；蹂躏，摧残



offshore *adj*. 离开海岸的；近海的；海外的，国外的
utility *n*. 实验程序；公用事业
segregate *vt*. （使）分开；分离；隔离
diode *n*. 二极管
configure *n*. 配置；设定；使成形；使具一定形式
salient *adj*. 显著的，突出的；重要的，主要的；跳跃的
segregation *n*. 分离，隔离
latency *n*. 潜伏；潜在因素
non optimized *adj*. 非优化的
enforceable *adj*. 可实施的；强行的；可强迫的
reconfiguration *n*. 重新配置，再组和；结构变形
falsify *vt*. 篡改，伪造
superfluous *adj*. 过多的；多余的；不必要的；奢侈的
stripped-down *adj*. 无装饰的；简装的
tactic *n*. 策略，战略
monopoly *n*. 垄断；垄断者；专卖权
insofar *adv*. 在……的范围；在……情况下
optimize *vt*. 使最优化，使完善
streamline *vt*. 把……做成流线型；组织；使合理化；使简单化
by-product *n*. 副产品；附带产生的结果；意外收获
fixation *n*. 固定；定位；定影
unto *prep*. 到，直到
sneak *adj*. 暗中进行的
fortuitous *adj*. 偶然的；不规则的

Exercises

Ⅰ. Answer the following questions.

1. What's the simple way to slim down the system?
2. Why hadn't the electric utilities gone through the same exercise as the oil companies?

Ⅱ. Translate the following sentences into Chinese.

1. General-purpose systems were being shared by multiple tasks, interleaving word processing and other routine work with critical functions that should have been relegated to specialized systems to prevent potential interference or disruption of operations.

____

____

____

____

2. The first salient aspect that surprised me from the meeting was that the people from the electric utilities and their electric utility oversight and clearinghouse organizations



did not refute my statements regarding the poor—or completely missing—security on their networks and systems.

____

____

____

____

3. One would be tempted to assume that deregulation and exposure of electric companies to a competitive environment would improve their efficiency and (following the same path as oil companies) their security.

____

____

____

____

Abstract Reading

An Information-theoretic Security Evaluation of a Class of Randomized Encryption Schemes

Randomized encryption techniques, where randomness is used for security enhancement, are considered. We focus on the case where the encrypted data experiences noise, e. g. , is transmitted over a noisy channel, within the encoding-encryption paradigm, where the data is first encoded for error correction, before being encrypted for security. We assume that the ciphertext is subject to a corruption equivalent to its transmission through a binary symmetric channel with known probability of error. The enhanced security is based on a dedicated wire-tap channel coding that introduces extra randomness, combined with that of the communication channel noise. The encryption is based on a block-by-block modulo 2 addition between an encoded message vector and a pseudorandom vector. The goal is to enhance the protection of the secret key employed in the encryption algorithm. Security evaluations of the model are performed employing an information-theoretic approach. Assuming both a passive and an active attacker, we show that there is a threshold before which the wire-tap encoder guarantees an information-theoretic security (during which the equivocation of the secret key is increased), and after which the uncertainty reduces, entering a regime in which a computational security analysis is needed for estimating the complexity resistance against the secret key recovery.



Adaptive Self-embedding Scheme With Controlled Reconstruction Performance

In this paper, we address the problem of adaptive self-embedding, where the reconstruction quality is controlled individually for different fragments of a digital image. We focus on the impact of incorporating content adaptivity features on the restoration success conditions and the achievable reconstruction performance. We analyze the problem theoretically and validate the obtained results experimentally with a fully functional selfembedding scheme. Our analysis shows that introduction of multiple reconstruction profiles, even with significantly lower restoration fidelity, does not need to improve the achievable tampering rate bounds. The obtained fine-grained control over the reconstruction process is exploited to provide guarantees on certain performance aspects. Based on the derived theoretical model, we propose a procedure for optimization of the overall reconstruction quality given constraints on the desired target tampering rate and the required quality level for selected image fragments. Such guarantees are of principal importance in a number of applications.

Design and Performance Analysis of a Virtual Ring Architecture for Smart Grid Privacy

The traditional electrical grid has become inadequate in meeting the needs and demands of electricity users in the 21st century. To address this challenge, smart grid technologies have emerged, which promise more efficient production and usage of electricity through bi-directional interactions between the consumer and the utility provider. This two-way interaction allows electricity to be generated in real time based on the actual needs of the consumers. However, this two-way interaction also raises concerns related to the privacy and the personal habits of consumers. To protect sensitive energy usage information of consumers, we propose a virtual ring architecture that can provide a privacy protection solution using symmetric or asymmetric encryptions of customers' requests belonging to the same group. We compare the efficiency of our proposed approach with two recently proposed smart grid privacy approaches namely, one based on blind signature and other based on a homomorphic encryption solution. We show that our approach maintains the privacy of customers while reducing the performance overhead of cryptographic computations by more than a factor of 2 when compared with the aforementioned past solutions. We further demonstrate that our smart grid privacy solution is simple, scalable, cost-effective, and incurs minimal computational processing overheads.

UNIT 7

COMPUTER SCIENCE

Text: Ethical Issues for Computer Scientists

Computers have had a powerful impact on our world and are destined to shape our future. This observation, now commonplace, is the starting point for any discussion of professionalism and ethics in computing.

The work of computer scientists and engineers is part of the social, political, economic, and cultural world in which we live, and it affects many aspects of that world. Professionals who work with computers have special knowledge. That knowledge, when combined with computers, has significant power to change people's lives—by changing sociotechnical systems; social, political and economic institutions; and social relationships.

In this paper, we provide a perspective on the role of computer and engineering professionals and we examine the relationships and responsibilities that go with having and using computing expertise. In addition to the topic of professional ethics, we briefly discuss several of the social-ethical issues created or exacerbated by the increasing power of computers and information technology: privacy, property, risk and reliability, and globalization.

Computers, digital data, and telecommunications have changed work, travel, education, business, entertainment, government, and manufacturing. For example, work now increasingly involves sitting in front of a computer screen and using a keyboard to make things happen in a manufacturing process or to keep track of records. In the past, these same tasks would have involved physically lifting, pushing, and twisting or using pens, paper, and file cabinets. Changes such as these in the way we do things have, in turn, fundamentally changed who we are as individuals, communities, and nations. Some would argue, for example, that new kinds of communities (e.g., cyberspace on the Internet) are forming, individuals are developing new types of personal identities, and new forms of au-



thority and control are taking hold as a result of this evolving technology.

Computer technology is shaped by social-cultural concepts, laws, the economy, and politics. These same concepts, laws, and institutions have been pressured, challenged, and modified by computer technology. Technological advances can antiquate laws, concepts, and traditions, compelling us to reinterpret and create new laws, concepts, and moral notions. Our attitudes about work and play, our values, and our laws and customs are deeply involved in technological change.

When it comes to the social-ethical issues surrounding computers, some have argued that the issues are not unique. All of the ethical issues raised by computer technology can, it is said, be classified and worked out using traditional moral concepts, distinctions, and theories. There is nothing new here in the sense that we can understand the new issues using traditional moral concepts, such as privacy, property, and responsibility, and traditional moral values, such as individual freedom, autonomy, accountability, and community. These concepts and values predate computers; hence, it would seem there is nothing unique about computer ethics.

On the other hand, those who argue for the uniqueness of the issues point to the fundamental ways in which computers have changed so many human activities, such as manufacturing, record keeping, banking, international trade, education, and communication. Taken together, these changes are so radical, it is claimed, that traditional moral concepts, distinctions, and theories, if not abandoned, must be significantly reinterpreted and extended. For example, they must be extended to computer-mediated relationships, computer software, computer art, datamining, virtual systems, and so on.

The uniqueness of the ethical issues surrounding computers can be argued in a variety of ways. Computer technology makes possible a scale of activities not possible before. This includes a larger scale of record keeping of personal information, as well as larger-scale calculations which, in turn, allow us to build and do things not possible before, such as undertaking space travel and operating a global communication system. Among other things, the increased scale means finer-grained personal information collection and more precise data matching and datamining. In addition to scale, computer technology has involved the creation of new kinds of entities for which no rules initially existed: entities such as computer files, computer programs, the Internet, Web browsers, cookies, and so on. The uniqueness argument can also be made in terms of the power and pervasiveness of computer technology. Computers and information technology seem to be bringing about a magnitude of change comparable to that which took place during the Industrial Revolution, transforming our social, economic, and political institutions; our understanding of what it means to be human; and the distribution of power in the world. Hence, it would seem that the is-



sues are at least special, if not unique.

In this paper we will take an approach that synthesizes these two views of computer ethics by assuming that the analysis of computer ethical issues involves both working on something new and drawing on something old. We will view issues in computer ethics as new species of older ethical problems, such that the issues can be understood using traditional moral concepts such as autonomy, privacy, property, and responsibility, while at the same time recognizing that these concepts may have to be extended to what is new and special about computers and the situations they create.

Most ethical issues arising around computers occur in contexts in which there are already social, ethical, and legal norms. In these contexts, often there are implicit, if not formal (legal), rules about how individuals are to behave; there are familiar practices, social meanings, interdependencies, and so on. In this respect, the issues are not new or unique, or at least cannot be resolved without understanding the prevailing context, meanings, and values. At the same time, the situation may have special features because of the involvement of computers—features that have not yet been addressed by prevailing norms. These features can make a moral difference. For example, although property rights and even intellectual property rights [1] had been worked out long before the creation of software, when software first appeared, it raised a new form of property issue. Should the arrangement of icons appearing on the screen of a user interface be ownable? Is there anything intrinsically wrong in copying software? Software has features that make the distinction between idea and expression (a distinction at the core of copyright law) almost incoherent.

As well, software has features that make standard intellectual property laws difficult to enforce. Hence, questions about what should be owned when it comes to software and how to evaluate violations of software ownership rights are not new in the sense that they are property rights issues, but they are new in the sense that nothing with the characteristics of software had been addressed before. We have, then, a new species of traditional property rights.

Similarly, although our understanding of rights and responsibilities in the employer-employee relationship has been evolving for centuries, never before have employers had the capacity to monitor their workers electronically, keeping track of every keystroke, and recording and reviewing all work done by an employee (covertly or with prior consent). When we evaluate this new monitoring capability and ask whether employers should use it, we are working on an issue that has never arisen before, although many other issues involving employer-employee rights have. We must address a new species of the tension between employer-employee rights and interests.



The social-ethical issues posed by computer technology are significant in their own right, but they are of special interest here because computer and engineering professionals bear responsibility for this technology. It is of critical importance that they understand the social change brought about by their work and the difficult social-ethical issues posed. Just as some have argued that the social-ethical issues posed by computer technology are not unique, some have argued that the issues of professional ethics surrounding computers are not unique. We propose, in parallel with our previous genus-species account, that the professional ethics issues arising for computer scientists and engineers are species of generic issues of professional ethics. All professionals have responsibilities to their employers, clients, co-professionals, and the public. Managing these types of responsibilities poses a challenge in all professions. Moreover, all professionals bear some responsibility for the impact of their work. In this sense, the professional ethics issues arising for computer scientists and engineers are generally similar to those in other professions. Nevertheless, it is also true to say that the issues arise in unique ways for computer scientists and engineers because of the special features of computer technology.

New Words and Phrases

destine *vt*. 注定
socio-technical *adj*. 社会技术的
exacerbate *vt*. 使加剧
file cabinet 文件柜
cyberspace *n*. (电子计算机创造的)通信、信息空间
antiquate *vt*. 使变成废弃物
predate *vt*. 提早日期，居先
reinterpret *vt*. 重新解释
data mining 数据挖掘
entity *n*. 实体；实际存在物；本质
pervasiveness *n*. 无处不在，遍布
magnitude *n*. 巨大，广大；重大，重要；量级
prevail *vi*. 流行，盛行；获胜，占优势；说服，劝说
intrinsically *adv*. 从本质上(讲)
incoherent *adj*. 不连贯的；不合逻辑的；无黏性的
keystroke *n*. 击键；按键
vt. 用键盘输入；击打……的键
covertly *adv*. 偷偷摸摸地；秘密地
consent *vi*. 同意；赞成；答应
n. 同意；(意见等的)一致；赞成
genus-species *n*. 种群

Notes

1. intellectual property rights：知识产权，指权利人对其所创作的智力劳动成果所享有的专有权利。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. evolving technology
2. professional ethics
3. digital data
4. personal identities
5. virtual system
6. global communication system
7. data matching
8. copyright law
9. monitoring capability

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

sequence	influential	priority	concepts
consequences	utilitarianism	access	analysis
trade-off	framework	minimized	policy
optimum	classified	objective	queue
ethical	dynamic	policies	elements

1. All of the ethical issues raised by computer technology can, it is said, be ________ and worked out using traditional moral ________, distinctions, and theories.

2. We can obtain a different performance ________ by implementing the sorted ________ by means of an array, which allows constant-time ________ to any element of the sequence given its position.

3. Realizing a ________ queue with a heap has the advantage that all of the operations take O(log N) time, where N is the number of ________ in the priority ________ at the time the operation is performed.

4. Our aim is not to propose, defend, or attack any particular ________ theory. Rather, we offer brief descriptions of three major and ________ ethical theories to illustrate the nature of ethical ________.

5. Utilitarianism has greatly influenced 20th-century thinking, especially insofar as it influenced the development of cost-benefit analysis. According to ________, we should

make decisions about what to do by focusing on the ________ of actions and policies; we should choose actions and ________ that bring about the best consequences.

6. The emphasis on consequences found in utilitarianism is very much a part of personal and policy decision making in our society, in particular as a ________ for law and public ________. Cost-benefit and risk-benefit analysis are, for example, consequentialist in character.

7. Optimization problems always have an ________ function to be ________ or maximized, but it is not often clear what steps to take to reach the ________ value. For example, in the optimum binary search tree problem of the previous section, we used ________ programming to systematically examine all possible trees.

Ⅲ. Translate the following sentences into Chinese.

1. Although usability test subjects normally escape actual bodily harm—even from irate developers resenting the users' mistreatment of their beloved software—test participation can still be quite distressing. Users feel a tremendous pressure to perform, even when told the study's purpose is to test the system and not the user.

____

____

____

____

2. Ethical analysis involves giving reasons for moral claims and commitments. It is not just a matter of articulating intuitions. When the reasons given for a claim are developed into a moral theory, the theory can be incorporated into techniques for improved technical decision making. The three ethical theories described in this section represent three traditions in ethical analysis and problem solving. The account we give is not exhaustive, nor is our description of the three theories any more than a brief introduction. The three traditions are utilitarianism, deontology, and social contract theory.

____

____

____

____

____

____

____

Reading: Ethical Issues That Arise from Computer Technology

The effects of a new technology on society can draw attention to an old issue and can change our understanding of that issue. The issues listed in this section—privacy, property rights, risk and reliability, and global communication—were of concern, even problematic, before computers were an important technology.

But computing and, more generally, electronic telecommunications, have added new-twists and new intensity to each of these issues. Although computer professionals cannot be expected to be experts on all of these issues, it is important for them to understand that computer technology is shaping the world.

And it is important for them to keep these impacts in mind as they work with computer technology. Those who are aware of privacy issues, for example, are more likely to take those issues into account when they design database management systems; those who are aware of risk and reliability issues are more likely to articulate these issues to clients and attend to them in design and documentation.

Privacy

Privacy is a central topic in computer ethics. Some have even suggested that privacy is a notion that has been antiquated by technology and that it should be replaced by a new openness. Others think that computers must be harnessed to help restore as much privacy as possible to our society. Although they may not like it, computer professionals are at the center of this controversy. Some are designers of the systems that facilitate information gathering and manipulation; others maintain and protect the information. As the saying goes, information is power —but power can be used or abused.

Computer technology creates wide-ranging possibilities for tracking and monitoring of human behavior. Consider just two ways in which personal privacy may be affected by computer technology. First, because of the capacity of computers, massive amounts of information can be gathered by record-keeping organizations such as banks, insurance companies, government agencies, and educational institutions. The information gathered can be kept and used indefinitely, and shared with other organizations rapidly and frequently. A second way in which computers have enhanced the possibilities for monitoring and tracking of individuals is by making possible new kinds of information. When activities are done using a computer, transactional information is created. When individuals use automated bank teller machines, records are created; when certain software is operating, keystrokes



on a computer keyboard are recorded; the content and destination of electronic mail can be tracked, and so on. With the assistance of newer technologies, much more of this transactional information is likely to be created. For example, television advertisers may be able to monitor television watchers with scanning devices that record who is sitting in a room facing the television. Highway systems allow drivers to pass through toll booths without stopping as a beam reading a bar code on the automobile charges the toll, simultaneously creating a record of individual travel patterns. All of this information (transactional and otherwise) can be brought together to create a detailed portrait of a person's life, a portrait that the individual may never see, although it is used by others to make decisions about the individual.

However, is it computer technology *per se* that poses the threat or is it just the way the technology has been used (and is likely to be used in the future)? Computer professionals might argue that they create the technology but are not responsible for how it is used.

This argument is, however, problematic for a number of reasons and perhaps foremost because it fails to recognize the potential for solving some of the problems of abuse in the design of the technology. Computer professionals are in the ideal position to think about the potential problems with computers and to design so as to avoid these problems. When, instead of deflecting concerns about privacy as out of their purview, computer professionals set their minds to solve privacy and security problems, the systems they design can improve.

At the same time we think about changing computer technology, we also must ask deeper questions about privacy itself and what it is that individuals need, want, or are entitled to when they express concerns about the loss of privacy. In this sense, computers and privacy issues are ethical issues. They compel us to ask deep questions about what makes for a good and just society. Should individuals have more choice about who has what information about them? What is the proper relationship between citizens and government, between individuals and private corporations? How are we to negotiate the tension between the competing needs for privacy and security? As previously suggested, the questions are not completely new, but some of the possibilities created by computers are new, and these possibilities do not readily fit the concepts and frameworks used in the past. Although we cannot expect computer professionals to be experts on the philosophical and political analysis of privacy, it seems clear that the more they know, the better the computer technology they produce is likely to be.

Property Rights and Computing

The protection of intellectual property rights has become an active legal and ethical de-

bate, involving national and international players. Should software be copyrighted, patented, or free? Is computer software a process, a creative work, a mathematical formalism, an idea, or some combination of these? What is society's stake in protecting software rights? What is society's stake in widely disseminating software? How do corporations and other institutions protect their rights to ideas developed by individuals? And what are the individuals' rights? Such questions must be answered publicly through legislation, through corporate policies, and with the advice of computing professionals. Some of the answers will involve technical details, and all should be informed by ethical analysis and debate.

An issue that has received a great deal of legal and public attention is the ownership of software. In the course of history, software is a relatively new entity. Whereas Western legal systems have developed property laws that encourage invention by granting certain rights to inventors, there are provisions against ownership of things that might interfere with the development of the technological arts and sciences. For this reason, copyrights protect only the expression of ideas, not the ideas themselves, and we do not grant patents on laws of nature, mathematical formulas, and abstract ideas. The problem with computer software is that it has not been clear that we could grant ownership of it without, in effect, granting ownership of numerical sequences or mental steps. Software can be copyrighted, because a copyright gives the holder ownership of the expression of the idea (not the idea itself), but this does not give software inventors as much protection as they need to compete fairly. Competitors may see the software, grasp the idea, and write a somewhat different program to do the same thing. The competitor can sell the software at less cost because the cost of developing the first software does not have to be paid. Patenting would provide stronger protection, but until quite recently the courts have been reluctant to grant this protection because of the problem previously mentioned: patents on software would appear to give the holder control of the building blocks of the technology, an ownership comparable to owning ideas themselves. In other words, too many patents may interfere with technological development.

Like the questions surrounding privacy, property rights in computer software also lead back to broader ethical and philosophical questions about what constitutes a just society. In computing, as in other areas of technology, we want a system of property rights that promotes invention (creativity, progress), but at the same time, we want a system that is fair in the sense that it rewards those who make significant contributions but does not give anyone so much control that others are prevented from creating. Policies with regard to property rights in computer software cannot be made without an understanding of the technology. This is why it is so important for computer professionals to be involved in public discussion and policy setting on this topic.



Risk, Reliability, and Accountability

As computer technology becomes more important to the way we live, its risks become more worrisome. System errors can lead to physical danger, sometimes catastrophic in scale. There are security risks due to hackers and crackers. Unreliable data and intentional misinformation are risks that are increased because of the technical and economic characteristics of digital data. Furthermore, the use of computer programs is, in a practical sense, inherently unreliable. Each of these issues (and many more) requires computer professionals to face the linked problems of risk, reliability, and accountability. Professionals must be candid about the risks of a particular application or system. Computing professionals should take the lead in educating customers and the public about what predictions we can and cannot make about software and hardware reliability. Computer professionals should make realistic assessments about costs and benefits, and be willing to take on both for projects in which they are involved.

There are also issues of sharing risks as well as resources. Should liability fall to the individual who buys software or to the corporation that developed it? Should society acknowledge the inherent risks in using rapidly evolving globally networked telecommunications?

The system of computers and connections known as the Internet provides the infrastructure for new kinds of communities—electronic communities. Questions of individual accountability and social control, as well as matters of etiquette, arise in electronic communities, as in all societies. It is not just that we have societies forming in a new physical environment; it is also that ongoing electronic communication changes the way individuals understand their identity, their values, and their plans for their lives. The changes that are taking place must be examined and understood, especially the changes affecting fundamental social values such as democracy, community, freedom, and peace.

Of course, speculating about the Internet is now a popular pastime, and it is important to separate the hype from the reality. The reality is generally much more complex and much more subtle. We will not engage in speculation and prediction about the future. Rather, we want to emphasize how much better off the world would be if (instead of watching social impacts of computer technology after the fact) computer engineers and scientists were thinking about the potential effects early in the design process. Of course, this can only happen if computer scientists and engineers are encouraged to see the social-ethical issues as a component of their professional responsibility.

New Words and Phrases

twist *vt*. 捻；拧使苦恼
　　n. 扭曲；拧；扭伤
articulate *vt*. 清晰地发(音)；明确有力地表达；用关节连接；使相互连贯
documentation *n*. 文件编制；文件的提供；文档编制
harness *vt*. 利用；控制
controversy *n*. 争论；争议；辩论
facilitate *vt*. 促进 ；使容易
toll booths 收费亭
philosophical *adj*. 哲学上的
stake *n*. 桩，赌注，奖金
　　vt. 资助，支持，把……押下打赌
legislation *n*. 立法；法律的制定；法规
provision *n*. 规定；条款；准备；供应品
　　vt. 供给……食物及必需品
patent *n*. 专利
　　vt. 获得……专利
catastrophic *adj*. 灾难的；惨重的，悲惨结局的
hacker *n*. 黑客
cracker *n*. 骇客
inherently *adv*. 天性地，本质地
infrastructure *n*. 基础设施
ongoing *adj*. 不间断的，进行的；前进的
　　n. 前进；行为，举止
speculation *n*. 推断 ；沉思

Exercises

Ⅰ. Answer the following questions.

1. What are the ethical issues that arise from computer technology?
2. Why are computers and privacy issues ethical issues?
3. What risks might computer technology bring about?

Ⅱ. Translate the following sentences into Chinese.

1. The effects of a new technology on society can draw attention to an old issue and can change our understanding of that issue. The issues listed in this section—privacy, property rights, risk and reliability, and global communication—were of concern, even problematic, before computers were an important technology.

2. Those who are aware of privacy issues, for example, are more likely to take those



issues into account when they design database management systems; those who are aware of risk and reliability issues are more likely to articulate these issues to clients and attend to them in design and documentation.

____

____

____

____

3. Although we cannot expect computer professionals to be experts on the philosophical and political analysis of privacy, it seems clear that the more they know, the better the computer technology they produce is likely to be.

____

____

____

____

Abstract Reading

Automatic Generation of Miniaturized Synthetic Proxies for Target Applications to Efficiently Design Multicore Processors

Prohibitive simulation time with pre-silicon design models and unavailability of proprietary target applications make microprocessor design very tedious. The framework proposed in this paper is the first attempt to automatically generate synthetic benchmark proxies for real world multithreaded applications. The framework includes metrics that characterize the behavior of the workloads in the shared caches, coherence logic, out-of-order cores, interconnection network and DRAM. The framework is evaluated by generating proxies for the workloads in the multithreaded PARSEC benchmark suite and validating their fidelity by comparing the microarchitecture dependent and independent metrics to that of the original workloads. The average error in IPC is 4.87 percent and maximum error is 10.8 percent for Raytrace in comparison to the original workloads. The average error in the power-per-cycle metric is 2.73 percent with a maximum of 5.5 percent when compared to original workloads. The representativeness of the proxies to that of the original workloads in terms of their sensitivity to design changes is evaluated by finding the correlation coefficient between the trends followed by the synthetic and the original for design changes in IPC, which is 0.92. A speedup of four to six orders of magnitude is achieved by using the synthetic proxies over the original workloads.

Coordinating Garbage Collection for Arrays of Solid-state Drives

Although solid-state drives (SSDs) offer significant performance improvements over hard disk drives (HDDs) for a number of workloads, they can exhibit substantial variance in request latency and throughput as a result of garbage collection (GC). When GC conflicts with an I/O stream, the stream can make no forward progress until the GC cycle completes. GC cycles are scheduled by logic internal to the SSD based on several factors such as the pattern, frequency, and volume of write requests. When SSDs are used in a RAID with currently available technology, the lack of coordination of the SSD-local GC cycles amplifies this performance variance. We propose a global garbage collection (GGC) mechanism to improve response times and reduce performance variability for a RAID of SSDs. We include a high-level design of SSD-aware RAID controller and GGC-capable SSD devices and algorithms to coordinate the GGC cycles. We develop reactive and proactive GC coordination algorithms and evaluate their I/O performance and block erase counts for various workloads. Our simulations show that GC coordination by a reactive scheme improves average response time and reduces performace variability for a wide variety of enterprise workloads. For bursty, write-dominated workloads, response time was improved by 69 percent and performance variability was reduced by 71 percent. We show that a proactive GC coordination algorithm can further improve the I/O response times by up to 9 percent and the performance variability by up to 15 percent. We also observe that it could increase the lifetimes of SSDs with some workloads (e. g. , Financial) by reducing the number of block erase counts by up to 79 percent relative to a reactive algorithm for write-dominant enterprise workloads.

Optimal Channel Estimation in Beamformed Systems for Common-randomness-based Secret Key Establishment

Establishing secret keys from the commonly-observed randomness of reciprocal wireless propagation channels has recently received considerable attention. We consider such key establishment between two multiantenna nodes that use beamforming for communication, showing that the upper bound on the number of key bits that can be generated from the channel observations can be maximized by properly probing the channel. Specifically, we demonstrate that the eigenvectors of the channel spatial covariance matrix should be used as beamformer weights during channel estimation and we optimize the energy allocated to channel estimation for each beamformer weight under a total energy constraint. Finally, by assuming that the channel covariance is separable, we illustrate implementation of the technique for practical beamforming systems and more advanced signal models that incorporate antenna mutual coupling.

UNIT 8

CRYPTOGRAPHY

Text: Public-key Infrastructures

1. Personal Security Environments

Importance

If Bob wants to generate signatures or decrypt documents using a public-key system, then he needs a private key. Bob must keep this key secret because everybody who knows the key can sign messages in Bob's name or decrypt secret documents that were sent to Bob. Therefore, Bob needs a personal security environment (PSE) in which his private keys are securely stored. Since the private keys should not leave the PSE, it also does the signing or decrypting.

Frequently, the PSE also generates the private keys. If the private keys are generated elsewhere, then at least the generating institution knows Bob's secret keys, which maycorrupt the security of the system. On the other hand, secure key generation may require resources not present in the PSE. For example, for RSA[1] keys random primes of a fixed bit length are required. In particular, the key generating environment must generate large, cryptographically secure, random numbers. If the random number generator of the PSE is weak, then the public-key system is insecure. It may therefore make sense to have the RSA keys generated by a trusted institution.

Implementation

The more sensitive the documents that are signed or encrypted, the more secure the

PSE must be. A simple PSE is a file in Bob's home directory that can be accessed only after entering a secret password.

This password may, for example, be used to decrypt the information. The security of a software PSE relies on the security of the underlying operating system. One may argue that operating systems must be very secure anyway and that they are therefore able to protect the PSE. Operating systems, for example, prevent unauthorized users from becoming administrators. On the other hand, it is well known that with sufficient effort the security of most operating systems can be successfully attacked. Therefore, a software PSE is not adequate for applications that require high security.

It is more secure to put the PSE on a smart card. Bob can carry his smart card in his wallet. If the card is in the smart card reader, it only permits very limited access. Manipulating its hardware or software is very difficult (although successful attacks have been reported).

Unfortunately, computations on smart cards are still very slow. Therefore, it is impossible to decrypt large documents on a smart card, so public keys encrypt session keys which, in turn, are used to encrypt the documents. The encrypted session key is appended to the encrypted document. The smart card only decrypts the session key. The decryption of the document is then done on a fast PC or workstation.

Representation Problem

Even if Alice uses a smart card for signing, there is still a severe security problem. If Alice wants to sign a document, she starts a program on her PC, which sends the document or its hash value to the smart card, where it is signed. With some effort, the attacker, Oscar, can manipulate the signing program on Alice's PC such that it sends a document to the smart card that is different from the one that Alice intended to sign. Because the smart card has no display, Alice is unable to detect this fraud. It is therefore possible that Alice could sign documents that she never wanted to sign. This problem is called the representation problem for signatures. The more important documents are for which digital signatures are accepted, the more dramatic the representation problem becomes. The problem is solved if Alice sees what she signs. For this purpose, Alice's PSE needs a display. One possibility is to use a cellular phone as a PSE. But its display is very small. Hence, the documents that can be signed securely on it are rather short. It depends on the solution of the representation problem whether digital signatures can be used to replace handwritten signatures.



2. Certification Authorities

If Alice uses a public-key system, it is not sufficient for her to keep her own private keys secret. If she uses the public key of Bob, she must be sure that it is really Bob's key. If the attacker, Oscar, is able to substitute his own public key for Bob's public key, then Oscar can decrypt secret messages to Bob and he can sign documents in Bob's name.

One solution of this problem is to establish trusted authorities. Each user is associated with such a certification authority(CA). The user trusts his CA. With its signature, the CA certifies the correctness and validity of the public keys of its users. The users know their CA's public key. Therefore, they can verify the signatures of their CA. We now explain in more detail what a CA does.

Registration

If Bob becomes a new user of the public-key system, then he is registered by his CA. He tells the CA his name and other relevant personal data. The CA verifies Bob's information. Bob can, for example, go to the CA in person and present some identification. The CA issues a user name for Bob that is different from the user name of all other users in the system. Bob ill use this name, for example, if he signs documents. If Bob wants to keep his name secret, then he may use a pseudonym. Then, only the CA knows Bob's real name.

Key Generation

Bob's public and private keys are generated either in his PSE or by his CA. It is recommended that Bob not know his private keys, because then he cannot inform others about those keys. The private keys are stored in Bob's PSE. The public keys are stored in a directory of the CA. Clearly, the keys must be protected while they are communicated between Bob and his CA.

For each purpose (for example, signing, encryption, and identification), a separate key pair is required. Otherwise, the system may become insecure. This is illustrated in the next example.

Certification

The CA generates a certificate, which establishes a verifiable connection between Bob and his public keys. This certificate is a string, which is signed by the CA and contains at least the following information.

1. The user name or the pseudonym of Bob.
2. Bob's public keys.
3. The names of the algorithms in which the public keys are used.
4. The serial number of the certificate.
5. The beginning and end of the validity of the certificate.
6. The name of the CA.
7. Restrictions that apply to the use of the certificate.

The certificate is stored, together with the user name, in a directory. Only the CA is allowed to write in this directory, but all users of the CA can read the information in the directory.

New Words and Phrases

cryptography *n.* 密码学
decrypt *vt.* 解码
corrupt *adj.* 错误百出的
vt. & *vi.* 损坏
prime *adj.* 素数的
append *vt.* 附加，添加
fraud *n.* 欺诈，伪劣品
validity *n.* 有效，正确，正确性
identification *n.* 识别试验，结构鉴定
pseudonym *n.* 假名，化名
certification *n.* 证明，鉴定，证书
directory *n.* 目录
adj. 指导的，咨询的，管理的

Notes

1. RSA：RSA 公钥加密算法是 1977 年由罗纳德·李维斯特(Ron Rivest)、阿迪·萨莫尔(Adi Shamir)和伦纳德·阿德曼(Leonard Adleman)一起提出的。当时他们三人都在麻省理工学院工作。RSA 就是他们三人姓氏开头字母拼在一起组成的。RSA 是目前最有影响力的公钥加密算法，它能够抵抗目前为止已知的绝大多数密码攻击，已被 ISO 推荐为公钥数据加密标准。只有短的 RSA 钥匙才可能被强力方式解破。到 2008 年为止，世界上还没有任何可靠的攻击 RSA 算法的方式。只要其钥匙的长度足够长，用 RSA 加密的信息实际上是不能被解破的。但在分布式计算和量子计算机理论日趋成熟的今天，RSA 加密安全性受到了挑战。RSA 算法基于一个十分简单的数论事实：将两个大素数相乘十分容易，但想要对其乘积进行因式分解却极其困难，因此可以将乘积公开作为加密密钥。



Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. private key
2. personal security environment
3. key generation
4. public-key system
5. smart card reader
6. session key
7. encrypted session
8. hash value
9. certification authority

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

certificates	authentication	public-key	exchanged
verified	keys	directory	validity
insecure	management	signature	encryption
distribution	replaced	secret	cryptosystems

1. Since public keys in asymmetric ________ need not be kept secret, key ________ in those systems is simpler than in symmetric schemes. Private keys, however, must be kept ________. Also, public ________ must be protected from falsification and abuse. Therefore, appropriate public-key infrastructures must be set up. They are responsible for key ________ and management.

2. Depending on their use, keys in ________ systems must be stored even after they expire. Public ________ keys must be stored as long as signatures generated with those keys must be ________. The CA stores ________ for public signature keys.

3. Authentication keys, private signature keys, and public ________ keys need not be put in archives. They must be stored only as long as they are used for ________, generating signatures, or encrypting documents.

4. The CA maintains a ________ of all certificates together with the name of the owner of each certificate.

5. All keys in a public-key system have a certain period of ________. Before a key ex-



pires, it must be ________ by a new, valid key. This new key is ________ between the CA and the users in such a way that it does not become ________ even if the old, invalid key becomes known.

Ⅲ. Translate the following paragraphs into Chinese.

1. In public-key infrastructures it is frequently useful to be able to reconstruct private keys. For example, if a user has lost his smart card that contains his private decryption key, then he cannot decrypt any encrypted file on his computer anymore. So those encrypted files are then inaccessible for the user unless it is possible to reconstruct the decryption key. However, for security reasons it may be important that the key cannot be reconstructed by a single person. That person could abuse the knowledge of the private key. It is more secure if a group of people has to be involved in the reconstruction.

____

____

____

____

____

____

____

2. The security of public-key cryptosystems is based on the intractability of certain computational problems. The security of the RSA and Rabin schemes is based on the hardness of integer factorization. The security of the ElGamal protocols and of DSA is based on the intractability of computing discrete logarithms in finite prime fields. However, none of those computational problems is provably intractable. Algorithmic progress has almost always been faster than predicted and it is known that quantum computers will make integer factorization and discrete logarithm computation in the relevant groups easy.

____

____

____

____

____

____

____





Reading: Passwords

Access to UNIX or Windows NT systems is typically controlled by password systems. Each user picks his individual and secret password w. The computer stores the image $f(w)$ of the password w under a one-way function f. If the user wants access to his computer, he enters his name and password w. The computer determines $f(w)$ and compares this value with the stored value. If they are identical, then access is granted. Otherwise, the user is rejected.

Passwords are also used to control access to World Wide Web pages or to files that contain private encryption or signature keys.

The password file does not need to be kept secret since it contains only the images $f(w)$ of the passwords w and f is a one-way function. Nevertheless, password identification systems are not very secure.

A user must memorize his or her password. Therefore, many users choose the first name of their spouse or children as their password. An attacker can mount a dictionary attack. For all words w in a dictionary, he computes $f(w)$ and compares the result with the entries in the password file. If he finds an entry of the password file, he has determined the corresponding password. It is, therefore, recommended to use symbols such as $ or # in the passwords. Then dictionary attacks are impossible, but it is also harder to memorize the passwords. It is also possible to store the password on a smart card. Instead of typing in his password, the prover inserts his smart card into the smart card reader. The verifier reads the password from the smart card. There is no need for the user to memorize or even know the password. On the contrary, if the user does not know his password he cannot give it away.

An attacker can also tap the connection between the prover and the verifier and can learn the password. This is particularly successful if there is a great distance between the prover and the verifier; for example, if a password system is used to protect World Wide Web access. Note that the use of smart cards does not prevent this attack.

Finally, the attacker can also replace an entry $f(w)$ in the password file with the image $f(v)$ of his own password v. Then, using the password v, he can get access. Therefore, the password file must be write protected.

One-Time Passwords

Using passwords is dangerous because an attacker can learn the passwords by tapping

the connection between the prover and the verifier. With one-time passwords, this attack does not work. One-time passwords are used for one identification. For the next identification, a new one-time password is used.

A simple way of implementing one-time passwords is the following. The verifier has a list $f(w_1)$, $f(w_2)$, $\cdots$, $f(w_n)$ of images of passwords w_1, $\cdots$, w_n. The prover knows this list of passwords and uses its elements for the identifications. Since the prover must store all passwords in advance, an attacker could learn some or all of them.

It is also possible that the prover and the verifier share a secret function f of an initial string w. Then the one-time passwords are $W_i = f^i(w)$, i :::: o. The prover can put the current password W_i and the one-way function f on a smart card. He does not need a large password file.

Challenge-Response Identification

Password identification system protocols have the disadvantage that an attacker can learn passwords long before the actual identification. This is even true for one-time password systems. Challenge response identification systems do not have this problem.

Alice wants to identify herself to Bob in a challenge response system. Bob asks a question, the challenge. Alice computes the response using her secret key and sends it to Bob. Bob verifies the response using the same secret key or the corresponding public key.

Symmetric Systems

We describe a simple challenge response identification system which uses a symmetric cryptosystem. We assume that the encryption key and the corresponding decryption key are the same. Alice and Bob share a secret key k. Alice wants to identify herself to Bob. Bob sends a random number r to Alice. Alice encrypts this random number by computing $c = E_k(r)$ and sends the ciphertext c to Bob. Bob decrypts the ciphertext; that is, he computes $r' = D_k(c)$ and compares the result with his chosen random number r. If $r = r'$, then he accepts the proof of identity; otherwise he rejects it.

This protocol proves that Alice knows the secret key at the time of the identification. It is not possible for Bob or an attacker to compute or obtain the correct response in advance. But since the verifier, Bob, also knows Alice's secret key, this key cannot be used for identification with another verifier since Bob can then pretend that he is Alice.

Public-key Systems

Challenge response systems can also be based on public-key signature schemes. If Alice wants to identify herself, she obtains a random number from Bob and signs this random



number with her private key. Bob verifies the signature, thereby verifying the identity of Alice.

In this protocol, Bob cannot pretend that he is Alice. He only knows Alice's public key. But it is necessary that Bob obtains the authentic public key of Alice. If the attacker, Oscar, can replace Alice's public key with his own, then he can convince Bob that he is Alice.

New Words and Phrases

encryption *n.* 加密；加密术
dictionary attack 字典攻击
verifier *n.* 检验者；核实者；校对机
protocol *n.* 协议；草案
symmetric *adj.* 对称的；匀称的
cryptosystem *n.* 密码系统
ciphertext *n.* 密文

Exercises

Ⅰ. Answer the following questions.

1. What does dictionary attack mean?
2. What's the simple way of implementing one-time passwords?
3. What's the disadvantage of password identification system protocols?

Ⅱ. Translate the following sentences into Chinese.

1. Access to UNIX or Windows NT systems is typically controlled by password systems. Each user picks his individual and secret password w. The computer stores the image $f(w)$ of the password w under a one-way function f. If the user wants access to his computer, he enters his name and password w. The computer determines $f(w)$ and compares this value with the stored value. If they are identical, then access is granted. Otherwise, the user is rejected.

______

2. An attacker can also tap the connection between the prover and the verifier and can

learn the password. This is particularly successful if there is a great distance between the prover and the verifier; for example, if a password system is used to protect World Wide Web access. Note that the use of smart cards does not prevent this attack.

3. Password identification system protocols have the disadvantage that an attacker can learn passwords long before the actual identification. This is even true for one-time password systems. Challenge response identification systems do not have this problem.

Abstract Reading

On-line Ciphers and the Hash-CBC Constructions

We initiate a study of on-line ciphers. These are ciphers that can take input plaintexts of large and varying lengths and will output the ith block of the ciphertext after having processed only the first i blocks of the plaintext. Such ciphers permit length-preserving encryption of a data stream with only a single pass through the data. We provide security definitions for this primitive and study its basic properties. We then provide attacks on some possible candidates, including CBC with fixed IV. We then provide two constructions, HCBC1 and HCBC2, based on a given block cipher E and a family of computationally AXU functions. HCBC1 is proven secure against chosen-plaintext attacks assuming that E is a PRP secure against chosen-plaintext attacks, while HCBC2 is proven secure against chosen-ciphertext attacks assuming that E is a PRP secure against chosen-ciphertext attacks.

Security Analysis of Randomize-hash-then-sign Digital Signatures

At CRYPTO 2006, Halevi and Krawczyk proposed two randomized hash function



modes and analyzed the security of digital signature algorithms based on these constructions. They showed that the security of signature schemes based on the two randomized hash function modes relies on properties similar to the second preimage resistance rather than on the collision resistance property of the hash functions. One of the randomized hash function modes was named the RMX hash function mode and was recommended for practical purposes. The National Institute of Standards and Technology (NIST), U. S. A. standardized a variant of the RMX hash function mode and published this standard in the Special Publication (SP) 800-106.

In this article, we first discuss a generic online birthday existential forgery attack of Dang and Perlner on the RMX-hash-then-sign schemes. We show that a variant of this attack can be applied to forge the other randomize-hash-then-sign schemes. We point out practical limitations of the generic forgery attack on the RMX-hash-then-sign schemes. We then show that these limitations can be overcome for the RMX-hash-then-sign schemes if it is easy to find fixed points for the underlying compression functions, such as for the Davies-Meyer construction used in the popular hash functions such as MD5 designed by Rivest and the SHA family of hash functions designed by the National Security Agency (NSA), USA and published by NIST in the Federal Information Processing Standards (FIPS). We show an online birthday forgery attack on this class of signatures by using a variant of Dean's method of finding fixed-point expandable messages for hash functions based on the Davies-Meyer construction. This forgery attack is also applicable to signature schemes based on the variant of RMX standardized by NIST in SP 800-106. We discuss some important applications of our attacks and discuss their applicability on signature schemes based on hash functions with "built-in" randomization. Finally, we compare our attacks on randomize-hash-then-sign schemes with the generic forgery attacks on the standard hash-based message authentication code (HMAC).

Security Models and Proof Strategies for Plaintext-aware Encryption

Plaintext-aware encryption is a simple concept: a public-key encryption scheme is plaintext aware if no polynomial-time algorithm can create a ciphertext without "knowing" the underlying message. However, the formal definitions of plaintext awareness are complex. This paper analyses these formal security definitions and presents the only known viable strategy for proving a scheme is PA2 plaintext aware. At the heart of this strategy is a new notion called PA1+ plaintext awareness. This security notion conceptually sits between PA1 and PA2 plaintext awareness (although it is formally distinct from either of these notions). We show exactly how this new security notion relates to the existing notions and how it can be used to prove PA2 plaintext awareness.

UNIT 9

INFORMATION RETRIEVAL

Text: Document Delineation and Character Sequence Decoding

Obtaining the Character Sequence in a Document

Digital documents that are the input to an indexing process are typically bytes in a file or on a web server. The first step of processing is to convert this byte sequence into a linear sequence of characters. For the case of plain English text in ASCII encoding, this is trivial. But often things get much more complex. The sequence of characters may be encoded by one of various single-byte or multibyte encoding schemes, such as Unicode UTF-8[1], or various national or vendor-specific standards. We need to determine the correct encoding. This can be regarded as a machine learning classification problem, but is often handled by heuristic methods, user selection, or using provided document metadata. Once the encoding is determined, we decode the byte sequence to a character sequence. We might save the choice of encoding because it gives some evidence about what language the document is written in.

The characters may have to be decoded out of some binary representation like Microsoft Word DOC files and/or a compressed format such as zip files. Again, we must determine the document format, and then an appropriate decoder has to be used. Even for plain text documents, additional decoding may need to be done. In XML[2] documents, character entities, such as &, need to be decoded to give the correct character, namely, & for &. Finally, the textual part of the document may need to be extracted out of other material that will not be processed. This might be the desired handling for XML files, if the markup is going to be ignored; we would almost certainly want to do this with postscript or PDF files. We do not deal further with these issues in this book, and assume



henceforth that our documents are a list of characters. Commercial products usually need to support a broad range of document types and encodings, because users want things to just work with their data as is. Often, they just think of documents as text inside applications and are not even aware of how it is encoded on disk. This problem is usually solved by licensing a software library that handles decoding document formats and character encodings.

The idea that text is a linear sequence of characters is also called into question by some writing systems, such as Arabic, where text takes on some two-dimensional and mixed-order characteristics, as shown in Figure 9.1 and Figure 9.2. But, despite some complicated writing system conventions, there is an underlying sequence of sounds being represented and hence an essentially linear structure remains. This is what is represented in the digital representation of Arabic, as shown in Figure 9.1.

كِتَابٌ ⇐ ك ِ ت َ ا ب ٌ

un b ā t i k

/kitābun/ 'a book'

Figure 9.1 An example of a vocalized Modern Standard Arabic word. The writing is from right to left and letters undergo complex mutations as they are combined. The representation of short vowels (here, /i/ and /u/) and the final /n/ (nunation) departs from strict linearity by being represented as diacritics above and below letters. Nevertheless, the represented text is still clearly a linear ordering of characters representing sounds. Full vocalization, as here, normally appears only in the Koran and children's books. Day-to-day text is unvocalized (short vowels are not represented, but the letter for ā would still appear) or partially vocalized, with short vowels inserted in places where the writer perceives ambiguities. These choices add further complexities to indexing.

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← → ← → ← START

Figure 9.2 The conceptual linear order of characters is not necessarily the order that you see on the page. In languages that are written right to left, such as Hebrew and Arabic, it is quite common to also have left-to-right text interspersed, such as numbers and dollar amounts. With modern Unicode representation concepts, the order of characters in files matches the conceptual order, and the reversal of displayed characters is handled by the rendering system, but this may not be true for documents in older encodings.

"Algeria achieved its independence in 1962 after 132 years of French occupation."

Choosing a Document Unit

The next phase is to determine what the document unit for indexing is. Thus unit far, we have assumed that documents are fixed units for the purposes of indexing. For example, we take each file in a folder as a document. But there are many cases in which you might want to do something different. A traditional UNIX (mbox-format) email file stores a sequence of email messages (an email folder) in one file, but you might wish to regard each email message as a separate document. Many email messages now contain attached documents, and you might then want to regard the email message and each contained attachment as separate documents. If an email message has an attached zip file, you might want to decode the zip file and regard each file it contains as a separate document. Going in the opposite direction, various pieces of web software (such as latex2html) take things that you might regard as a single document (e. g. , a PowerPoint file or a LATEX document) and split them into separate HTML pages for each slide or subsection, stored as separate files. In these cases, you might want to combine multiple files into a single document.

More generally, for very long documents, the issue of indexing granularity arises. For a collection of books, it would usually be a bad idea to index an entire book as a document. A search for Chinese toys might bring up a book that mentions China in the first chapter and toys in the last chapter, but this does not make it relevant to the query. Instead, we may well wish to index each chapter or paragraph as a mini-document. Matches are then more likely to be relevant, and because the documents are smaller it will be much easier for the user to find the relevant passages in the document. But why stop there? We could treat individual sentences as mini-documents. It becomes clear that there is a precision/recall tradeoff here. If the units get too small, we are likely to miss important passages because terms were distributed over several mini-documents, whereas if units are too large we tend to get spurious matches and the relevant information is hard for the user to find. The problems with large document units can be alleviated by use of explicit or implicit proximity search and the tradeoffs in resulting system performance that we are hinting at are discussed in later parts. The issue of index granularity, and in particular a need to simultaneously index documents at multiple levels of granularity, appears prominently in XML retrieval. An information retrieval (IR) system should be designed to offer choices of granularity. For this choice to be made well, the person who is deploying the system must have a good understanding of the document collection, the users, and their likely information needs and usage patterns. For now, we assume that a suitable size document unit has been chosen, together with an appropriate way of dividing or aggregating files, if needed.



New Words and Phrases

retrieval *n.* 检索
delineation *n.* 描述，描写
byte *n.* 字节
trivial *adj.* 琐碎的，无价值的
heuristic *adj.* 启发式的，探索的
metadata *n.* 元数据
entity *n.* 实体，本质
license *vt.* 同意，发许可证
vocalize *vt.* & *vi.* 成为有声，以声音表示出
mutation *n.* 变化，转变
vowel *n.* 元音，母音
nunation *n.* 词尾加“n”(如阿拉伯语某些名词的词尾变化)
linearity *n.* 线性
diacritic *adj.* 可区别的，读音符号的
perceive *vt.* & *vi.* 意识到，觉察，理解
ambiguity *n.* 含糊，歧义，模棱两可
Hebrew *n.* 希伯来人，犹太人，希伯来语
intersperse *vt.* 散布，散置
Unicode *n.* 统一码(采用双字节对字符进行编码)
folder *n.* 文件夹
subsection *n.* 小节
granularity *n.* 间隔尺寸，粒度
explicit *adj.* 明确的，清楚的
implicit *adj.* 无疑的，绝对的
proximity *n.* 接近，邻近；接近度
prominently *adv.* 显著地，重要地
aggregate *vt.* & *vi.* 总计达；使聚集

Notes

1. UTF-8：8-bit Unicode Transformation Format，是一种针对 Unicode 的可变长度字符编码，又称万国码。由 Ken Thompson 于 1992 年创建。现在已经标准化为 RFC 3629。UTF-8 用 1 到 4 个字节编码 Unicode 字符。用在网页上可以同一页面显示中文简体、繁体及其他语言(如日文、韩文)。

2. XML：可扩展标记语言，标准通用标记语言的子集，一种用于标记电子文件使其具有结构性的标记语言。它可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。它非常适合万维网传输，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. information retrieval
2. character sequence

3. encoding scheme
4. vendor-specific standard
5. heuristic methods
6. binary representation
7. HTML
8. aggregating file
9. linear sequence
10. document format

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

punctuation	sequence	corresponding	specified
strategies	collections	additionally	hashing
inefficient	query	merging	contain
inverted	documents	phrases	material
semantically	throwing	unstructured	tokens

1. Information retrieval (IR) is finding ________ (usually documents) of an ________ nature (usually text) that satisfies an information need from within large ________ (usually stored on computers).

2. Information retrieval can also cover other kinds of data and information problems beyond that ________ in the core definition above. The term "unstructured data" refers to data that does not have clear, ________ overt, easy-for-a-computer structure.

3. Given a character ________ and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time ________ away certain characters, such as ________.

4. Having broken up our documents (and also our query) into ________, the easy case is if tokens in the ________ just match tokens in the token list of the document. However, there are many cases when two character sequences are not quite the same but you would like a match to occur. For instance, if you search for USA, you might hope to also match ________ containing U. S. A.

5. For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. ________, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that ________ another word in the set.



6. The ________ of biword indexes and positional indexes can be fruitfully combined. If users commonly query on particular ________, such as Michael Jackson, it is quite ________ to keep ________ positional postings lists. A combination strategy uses a phrase index, or just a biword index, for certain queries and uses a positional index for other phrase queries.

7. Given an ________ index and a query, our first task is to determine whether each query term exists in the vocabulary and, if so, identify the pointer to the ________ postings. This vocabulary lookup operation uses a classical data structure called the dictionary and has two broad classes of solutions: ________ and search trees.

Ⅱ. Translate the following sentences into Chinese.

1. Our final technique for tolerant retrieval has to do with phonetic correction: misspellings that arise because the user types a query that sounds like the target term. Such algorithms are especially applicable to searches on the names of people. The main idea here is to generate, for each term, a "phonetic hash" so that similar-sounding terms hash to the same value.

2. One benefit of compression is immediately clear. We need less disk space. As we will see, compression ratios of 1∶4 are easy to achieve, potentially cutting the cost of storing the index by 75%.

3. There are two more subtle benefits of compression. The first is increased use of caching. Search systems use some parts of the dictionary and the index much more than others. The second more subtle advantage of compression is faster transfer of data from disk to memory. Efficient decompression algorithms run so fast on modern hardware that the total time of transferring a compressed chunk of data from disk and then decompressing it is usually less than transferring the same chunk of data in uncompressed form.

Reading: Search Structures for Dictionaries

Given an inverted index and a query, our first task is to determine whether each query term exists in the vocabulary and, if so, identify the pointer to the corresponding postings. This vocabulary lookup operation uses a classical data structure called the dictionary and has two broad classes of solutions: hashing and search trees. In the literature of data structures, the entries in the vocabulary (in our case, terms) are often referred to as keys. The choice of solution (hashing or search trees) is governed by a number of questions.

(1) How many keys are we likely to have? (2) Is the number likely to remain static, or change a lot? And in the case of changes, are we likely to only have new keys inserted, or to also have some keys in the dictionary be deleted? (3) What are the relative frequencies with which various keys will be accessed?

Hashing has been used for dictionary lookup in some search engines. Each vocabulary term (key) is hashed into an integer over a large enough space that hash collisions are unlikely; collisions are resolved by auxiliary structures that can demand care to maintain. At query time, we hash each query term separately and, following a pointer to the corresponding postings, taking into account any logic for resolving hash collisions. There is no easy way to find minor variants of a query term (such as the accented and unaccented versions of a word like resume), because these could be hashed to very different integers. In particular, we cannot seek (for instance) all terms beginning with the prefix automat. Finally, in a setting (such as the Web), where the size of the vocabulary keeps growing, a hash function designed for current needs may not suffice in a few years' time.

Search trees overcome many of these issues-for instance, they permit us to enumerate all vocabulary terms beginning with automat. The best-known search tree is the binary tree, in which each internal node has two children. The search for a term begins at the root of the tree. Each internal node (including the root) represents a binary test, based on whose outcome the search proceeds to one of the two subtrees below that node. Figure 9.3 gives an example of a binary search tree used for a dictionary. Efficient search (with a number of comparisons that is $O(\log M)$) hinges on the tree being balanced: the numbers of terms under the two subtrees of any node are either equal or differ by 1. The principal is-



sue here is that of rebalancing; as terms are inserted into or deleted from the binary search tree, it needs to be rebalanced so that the balance property is maintained.

To mitigate rebalancing, one approach is to allow the number of subtrees under an internal node to vary in a fixed interval. A search tree commonly used for a dictionary is the B-tree—a search tree in which every internal node has a number of children in the interval $[a, b]$, where a and

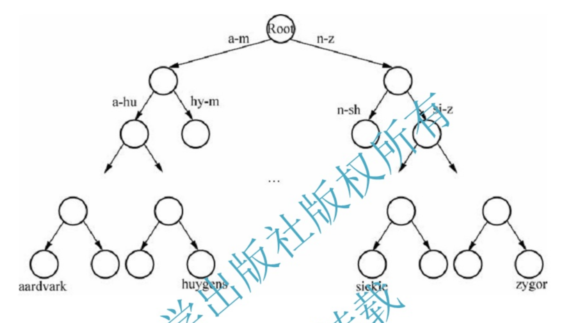


Figure 9.3 A binary search tree. In this example, the branch at the root partitions vocabulary terms into two subtrees, those whose first letter is between a and m, and the rest.

b are appropriate positive integers; Figure 9.4 shows an example with $a=2$ and $b=4$. Each branch under an internal node again represents a test for a range of character sequences, as in the binary tree example of Figure 9.3. A B-tree may be viewed as "collapsing" multiple levels of the binary tree into one; this is especially advantageous when some of the dictionary is disk resident, in which case this collapsing serves the function of pre-fetching imminent binary tests. In such cases, the integers a and b are determined by the sizes of disk blocks.

It should be noted that, unlike hashing, search trees demand that the characters used in the document collection have a prescribed ordering; for instance, the 26 letters of the English alphabet are always listed in the specific order A through Z. Some Asian languages such as Chinese do not always have a unique ordering, although by now all languages (including Chinese and Japanese) have adopted a standard ordering system for their character sets. Wildcard queries are used in any of the following situations: (1) the user is uncertain of the spelling of a query term (e.g., Sydney vs. Sidney, which leads to the wildcard query Sdney); (2) the user is aware of multiple variants of spelling a term and (consciously)

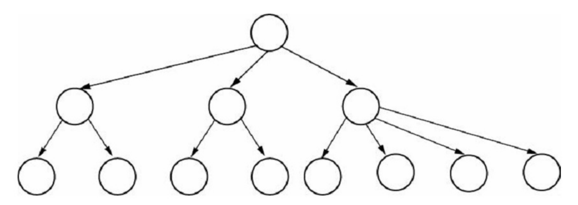


Figure 9.4 A B-tree. In this example every internal node has between 2 and 4 children.

seeks documents containing any of the variants (e.g., color vs. colour); (3) the user seeks documents containing variants of a term that would be caught by stemming, but is unsure whether the search engine performs stemming (e.g., judicial vs. judiciary, leading to the wildcard query judicia); or (4) the user is uncertain of the correct rendition of a foreign word or phrase (e.g., the query Universit Stuttgart).

A query such as mon is known as a trailing wildcard query, because the symbol occurs only once, at the end of the search string. A search tree on the dictionary is a convenient way of handling trailing wildcard queries: we walk down the tree following the symbols m, o, and n in turn, at which point we can enumerate the set W of terms in the dictionary with the prefix mon.

Finally, we use |W| lookups on the standard inverted index to retrieve all documents containing any term in W. But what about wildcard queries in which the * symbol is not constrained to be at the end of the search string? Before handling this general case, we mention a slight generalization of trailing wildcard queries. First, consider leading wildcard queries, or queries of the form *mon. Consider a reverse B-tree on the dictionary—one in which each root-to-leaf path of the B-tree corresponds to a term in the dictionary written backwards: thus, the term lemon would, in the B-tree, be represented by the pathroot-n-o-m-e-l. A walk down the reverse B-tree then enumerates all terms R in the vocabulary with a given prefix.

In fact, using a regular B-tree together with a reverse B-tree, we can handle an even more general case: wildcard queries in which there is a single symbol, such as semon. To do this, we use the regular B-tree to enumerate the set W of dictionary terms beginning with the prefix se and a non-empty suffix, then the reverse B-tree to enumerate the set R of terms ending with the suffix mon. Next, we take the intersection W ∩ R of these two sets, to arrive at the set of terms that begin with the prefix se and end with the suffix mon. Finally, we use the standard inverted index to retrieve all documents containing any terms in this intersection. We can thus handle wildcard queries that contain a single symbol using



two B-trees, the normal B-tree and a reverse B-tree.

New Words and Phrases

inverted *adj*. 反向的，倒转的
hashing *n*. 散列法
lookup *n*. 查找
collision *n*. 碰撞，冲突
variant *n*. (词等的)变体；变形
adj. 不同的，相异的
accented *adj*. 带……腔调的；带……口音的
enumerate *vt*. 列举，枚举，数
automat *n*. 自动机
subtree *n*. 子树，树状子目录
imminent *adj*. 即将发生的，迫切的
prescribe *vt*. & *vi*. 指定，规定
rendition *n*. 演奏；翻译；给予
trailing *adj*. 拖尾的；蔓延的
wildcard *n*. 通配符
constrain *vt*. 限制；约束

Exercises

Ⅰ. Answer the following questions.

1. What does hashing do for dictionary lookup in some search engines?
2. What's the approach to mitigate rebalancing?
3. What is efficient searching and what are its criteria?

Ⅱ. Translate the following sentences into Chinese.

1. Given an inverted index and a query, our first task is to determine whether each query term exists in the vocabulary and, if so, identify the pointer to the corresponding postings. This vocabulary lookup operation uses a classical data structure called the dictionary and has two broad classes of solutions: hashing and search trees.

____

____

____

____

____

2. Hashing has been used for dictionary lookup in some search engines. Each vocabulary term (key) is hashed into an integer over a large enough space that hash collisions are unlikely; collisions are resolved by auxiliary structures that can demand care to maintain.

____

3. It should be noted that, unlike hashing, search trees demand that the characters used in the document collection have a prescribed ordering; for instance, the 26 letters of the English alphabet are always listed in the specific order A through Z. Some Asian languages such as Chinese do not always have a unique ordering, although by now all languages (including Chinese and Japanese) have adopted a standard ordering system for their character sets.

Abstract Reading

Riding the Multimedia Big Data Wave

In this talk we present a perspective across multiple industry problems, including safety and security, medical, Web, social and mobile media, and motivate the need for large-scale analysis and retrieval of multimedia data. We describe a multi-layer architecture that incorporates capabilities for audio-visual feature extraction, machine learning and semantic modeling and provides a powerful framework for learning and classifying contents of multimedia data. We discuss the role semantic ontologies for representing audio-visual concepts and relationships, which are essential for training semantic classifiers. We discuss the importance of using faceted classification schemes in particular for organizing multimedia semantic concepts in order to achieve effective learning and retrieval. We also show how training and scoring of multimedia semantics can be implemented on big data distributed computing platforms to address both massive-scale analysis and low-latency processing. We describe multiple efforts at IBM on image and video analysis and retrieval, including IBM Multimedia Analysis and Retrieval System (IMARS), and show recent results for semantic-based classification and retrieval. We conclude with future directions for improving a-



nalysis of multimedia through interactive and curriculum-based techniques for multimedia semantics-based learning and retrieval.

Improving Search Result Summaries by Using Searcher Behavior Data

Query-biased search result summaries, or "snippets", help users decide whether a result is relevant for their information need, and have become increasingly important for helping searchers with difficult or ambiguous search tasks. Previously published snippet generation algorithms have been primarily based on selecting document fragments most similar to the query, which does not take into account which parts of the document the searchers actually found useful. We present a new approach to improving result summaries by incorporating post-click searcher behavior data, such as mouse cursor movements and scrolling over the result documents. To achieve this aim, we develop a method for collecting behavioral data with precise association between searcher intent, document examination behavior, and the corresponding document fragments. In turn, this allows us to incorporate page examination behavior signals into a novel Behavior-Biased Snippet generation system (BeBS). By mining searcher examination data, BeBS infers document fragments of most interest to users, and combines this evidence with text-based features to select the most promising fragments for inclusion in the result summary. Our extensive experiments and analysis demonstrate that our method improves the quality of result summaries compared to existing state-of-the-art methods. We believe that this work opens a new direction for improving search result presentation, and we make available the code and the search behavior data used in this study to encourage further research in this area.

Modeling and Analysis of Cross-session Search Tasks

The information needs of search engine users vary in complexity, depending on the task they are trying to accomplish. Some simple needs can be satisfied with a single query, whereas others require a series of queries issued over a longer period of time. While search engines effectively satisfy many simple needs, searchers receive little support when their information needs span session boundaries. In this work, we propose methods for modeling and analyzing user search behavior that extends over multiple search sessions. We focus on two problems: (i) given a user query, identify all of the related queries from previous sessions that the same user has issued, and (ii) given a multi-query task for a user, predict whether the user will return to this task in the future. We model both problems within a classification framework that uses features of individual queries and long-term user search

behavior at different granularity. Experimental evaluation of the proposed models for both tasks indicates that it is possible to effectively model and analyze cross-session search behavior. Our findings have implications for improving search for complex information needs and designing search engine features to support cross-session search tasks.

UNIT 10

IMAGE PROCESSING

Text: Basic Gray-level Image Processing

In this paper, we study basic gray-level digital image[1] processing operations. The types of operations studied fall into three classes.

The first are *point operations*, or image processing operations that are applied to individual pixels only. Thus, interactions and dependencies between neighboring pixels are not considered, nor are operations that consider multiple pixels simultaneously to determine an output. Since spatial information, such as a pixel's location and the values of its neighbors, are not considered, point operations are defined as functions of pixel intensity only. The basic tool for understanding, analyzing, and designing image point operations is the *image histogram*, which will be introduced below.

The second class includes *arithmetic operations* between images of the same spatial dimensions. These are also point operations in the sense that spatial information is not considered, although information is shared between images on a pointwise basis. Generally, these have special purposes, e.g., for noise reduction and change or motion detection.

The third class of operations are *geometric image operations*. These are complementary to point operations in the sense that they are not defined as functions of image intensity. Instead, they are functions of spatial position only. Operations of this type change the appearance of images by changing the coordinates of the intensities. This can be as simple as image translation or rotation, or it may include more complex operations that distort or bend an image, or "morph" a video sequence. Since our goal, however, is to concentrate on digital image processing of real-world images, rather than the production of special effects, only the most basic geometric transformations will be considered. More complex and time-varying geometric effects are more properly considered within the science of

computer graphics.

Notation

Point operations, algebraic operations, and geometric operations are easily defined on images of any dimensionality including digital video data. For simplicity of presentation, we will restrict our discussion to two-dimensional images only. The extensions to three or higher dimensions are not difficult, especially in the case of point operations, which are independent of dimensionality. In fact, spatial temporal information is not considered in their definition or application.

We will also only consider monochromatic images, since extensions to color or other multispectral images is either trivial, in that the same operations are applied identically to each band (e. g., R, G, B), or they are defined as more complex color space operations, which goes beyond what we want to cover in this basic paper.

Suppose then that the single-valued image $f(n)$ to be considered is defined on a two-dimensional discrete-space coordinate system $n = (n_1, n_2)$. The image is assumed to be of finite support, with image domain $[0, N-1] \times [0, M-1]$. Hence the nonzero image data can be contained in a matrix or array of dimensions $N \times M$ (rows, columns). This discrete-space image will have originated by sampling a continuous image $f(x, y)$. Furthermore, the image $f(n)$ is assumed to be quantized to K levels $\{0, \cdots, K-1\}$; hence each pixel value takes one of these integer values. For simplicity, we will refer to these values as gray levels, reflecting the way in which monochromatic images are usually displayed. Since $f(n)$ is both discrete-space and quantized, it is digital.

Image Histogram

The basic tool that is used in designing point operations on digital images (and many other operations as well) is the image histogram. The histogram H_f of the digital image f is a plot or graph of the *frequency of occurrence* of each gray level in f. Hence, H_f is a one-dimensional function with domain $(0, \cdots, K-1)$ and possible range extending from 0 to the number of pixels in the image, NM.

The histogram is given explicitly by

$$H_f(k) = J \tag{10-1}$$

if f contains exactly J occurrences of gray level k, for each $k = 0, \cdots, K-1$. Thus, an algorithm to compute the image histogram involves a simple counting of gray levels, which can be accomplished even as the image is scanned. Every image processing develop-



ment environment and software library contains basic histogram computation, manipulation, and display routines.

Since the histogram represents a reduction of dimensionality relative to the original image f, information is lost-the image f cannot be deduced from the histogram H_f except in trivial cases (when the image is constant valued). In fact, the number of images that share the same arbitrary histogram H_f is astronomical. Given an image f with a particular histogram H_f, every image that is a spatial shuffling of the gray levels of f has the same histogram H_f.

The histogram H_f contains no spatial information about f—it describes the frequency of the gray levels in f and nothing more. However, this information is still very rich, and many useful image processing operations can be derived from the image histogram. Indeed, a simple visual display of H_f reveals much about the image. By examining the appearance of a histogram, it is possible to ascertain whether the gray levels are distributed primarily at lower (darker) gray levels, or vice versa. Although this can be ascertained to some degree by visual examination of the image itself, the human eye has a tremendous ability to adapt to overall changes in luminance, which may obscure shifts in the gray-level distribution. The histogram supplies an absolute method of determining an image's gray-level distribution.

For example, the average optical density, or AOD, is the basic measure of an image's overall average brightness or gray level. It can be computed directly from the image:

$$\mathrm{AOD}(f) = \frac{1}{MN}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{M-1} f(n_1, n_2) \tag{10-2}$$

or it can be computed from the image histogram:

$$\mathrm{AOD}(f) = \frac{1}{MN}\sum_{k=0}^{K-1} kH_f(k) \tag{10-3}$$

The AOD is a useful and simple meter for estimating the center of an image's gray-level distribution. A target value for the AOD might be specified when designing a point operation to change the overall gray-level distribution of an image.

Figure1 depicts two hypothetical image histograms. The one on the left has a heavier distribution of gray levels close to zero (and a low AOD), while the one on the right is skewed toward the right (a high AOD). Since image gray levels are usually displayed with lower numbers' indicating darker pixels, the image on the left corresponds to a predominantly dark image. This may occur if the image f was originally underexposed prior to digitization, or if it was taken under poor lighting levels, or perhaps the process of digitization was performed improperly. A skewed histogram often indicates a problem in gray-level allocation. The image on the right may have been overexposed or taken in very bright light.

New Words and Phrases

spatial *adj*. 空间的，立体的
rotation *n*. 旋转，转动
distort *vt*. & *vi*. 扭曲，失真
morph *vt*. 变形
pointwise *adj*. 逐点的
complementary *adj*. 互补的；补充的
geometric *adj*. 几何的
spatialtemporal *n*. 时空
monochromatic *adj*. 单色的，单频的
multispectral *adj*. 多谱线的，多光谱的
finite *adj*. 有限的
    n. 有限性；有限的事物
nonzero *n*. 非零
matrix *n*. 矩阵
quantize *vt*. 量化
histogram *n*. 直方图
occurrence *n*. 发生，出现
astronomical *adj*. 天文学的，极大的
ascertain *vt*. 弄清，确定
tremendous *adj*. 极大的，惊人的
obscure *adj*. 晦涩的，不清楚的
depict *vt*. 描绘，描画
skew *n*. 歪曲；曲解
underexpose *vt*. 使(底片)曝光不足
overexpose *vt*. 曝光过度

Notes

1. gray-level digital image：在计算机领域中，灰度数字图像是每个像素只有一个采样颜色的图像。这类图像通常显示为从最暗的黑色到最亮的白色的灰度(黑色灰度为0，白色亮度为255)，尽管理论上这个采样可以是任何颜色的不同深浅，甚至可以是不同亮度上的不同颜色。灰度图像与黑白图像不同，在计算机图像领域中黑白图像只有黑白两种颜色，灰度图像在黑色与白色之间还有许多级的颜色深度。但是，在数字图像领域之外，“黑白图像”也表示“灰度图像”，例如，灰度的照片通常称为“黑白照片”。在一些关于数字图像的文章中单色图像等同于灰度图像，在另外一些文章中又等同于黑白图像。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. point operation
2. geometric operation
3. arithmetic operation
4. image histogram
5. motion detection



6. noise reduction
7. geometric image operation
8. time-varying
9. algebraic operation
10. discrete-space image
11. frequency of occurrence

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

gross	zoomed	geometric	shape
gray	multiple	representations	visual
transitions	redundancies	coordinate	surveillance
transform	boundary	statistics	equalization
separation	edges	applications	scene
magnifications	naive	pixel	motion

1. In some ________, it is desired to ________ the image into one that has a histogram of a specific ________. The process of histogram shaping generalizes histogram ________, which is the special case in which the target shape is flat. Histogram shaping can be applied when ________ images of the same scene, but taken under mildly different lighting conditions, are to be compared.

2. Detected ________ is also useful for tracking targets, for recognizing objects by their motion, and for computing three-dimensional ________ information from two-dimensional motion.

3. If the time ________ between frames is not small, then change detection can involve the discovery of ________ scene changes. This can be useful for security or ________ cameras, or in automated ________ inspection systems, for example.

4. The image zoom is a good example of a ________ operation for which the type of interpolation is important, particularly at high ________. With nearest neighbor interpolation, many values in the ________ image may be assigned the same ________ scale, resulting in a severe "blotching" or "blocking" effect.

5. At first glance, ________ detection may seem trivial, since the boundary points can be simply defined as the ________ from 1 to 0 (and vice versa). However, when there is noise present, boundary detection becomes quite sensitive to small noise artifacts, leading to many useless detected ________.

6. Compressed images are ________ that require less storage than the nominal stor-

age. This is generally accomplished by coding of the data based on measured ________, rearrangement of the data to exploit patterns and ________ in the data, and (in the case of lossy compression), quantization of information.

7. The basic idea of chain coding is to code contour directions instead of ________ bit-by-bit binary image coding or even ________ representations of the contours. Chain coding is based on identifying and storing the directions from each ________ to its neighbor pixel on each contour.

Ⅲ. Translate the following sentences into Chinese.

1. Images that occur in practical applications invariably suffer from random degradations that are collectively referred to as noise. These degradations arise from numerous sources, including radiation scatter from the surface before the image is sensed; electrical noise in the sensor or camera; channel noise as the image is transmitted over a communication channel; bit errors after the image is digitized, and so on.

2. Geometric image operations are, in a sense, the opposite of point operations: they modify the spatial positions and spatial relationships of pixels, but they do not modify graylevel values. Generally, these operations can be quite complex and computationally intensive, especially when applied to video sequences. However, the more complex geometric operations are not much used in engineering image processing, although they are heavily used in the computer graphics field.

3. A simple but powerful tool for identifying and labeling the various objects in a binary image is a process called region labeling, blob coloring, or connected component identification. It is useful since once they are individually labeled, the objects can be separately manipulated, displayed, or modified.



Reading: Basic Binary Image Processing

In this paper on basic methods, we explain and demonstrate fundamental tools for the processing of binary digital images. Binary image processing is of special interest, since an image in binary format can be processed with very fast logical (Boolean) operators. Often, a binary image has been obtained by abstracting essential information from a gray-level image, such as object location, object boundaries, or the presence or absence of some image property.

As we know, a digital image is an array of numbers or sampled image intensities. Each gray level is quantized or assigned one of a finite set of numbers represented by B bits. In a binary image, only one bit is assigned to each pixel: $B=1$, implying two possible gray-level values, 0 and 1. These two values are usually interpreted as Boolean; hence each pixel can take on the logical values 0 or 1, or equivalently, "true" or "false". For example, these values might indicate the absence or presence of some image property in an associated gray-level image of the same size, where 1 at a given coordinate indicates the presence of the property at that coordinate in the gray level image, and 0 otherwise. This image property is quite commonly a sufficiently high or low intensity (brightness), although more abstract properties, such as the presence or absence of certain objects, or smoothness or non-smoothness, etc., might be indicated.

Since most image display systems and software assume images of eight or more bits per pixel, the question arises as to how binary images are displayed. Usually, they are displayed using the two extreme gray tones, black and white, which are ordinarily represented by 0 and 255, respectively, in a gray-scale display environment. There is no established convention for the Boolean values that are assigned to "black" and to "white". In this paper we will uniformly use 1 to represent black (displayed as gray-level 0) and 0 to represent white (displayed as gray-level 255). However, the assignments are quite commonly reversed, and it is important to note that the Boolean values 0 and 1 have no physical significance other than what the user assigns to them.

Binary images arise in a number of ways. Usually, they are created from gray-level images for simplified processing or for printing. However, certain types of sensors directly deliver a binary image output. Such devices are usually associated with printed, handwrit-

ten, or line drawing images, with the input signal being entered by hand on a pressure sensitive tablet, a resistive pad, or a light pen. In such a device, the (binary) image is first initialized prior to image acquisition:

$$g(n) = 0 \tag{10-4}$$

at all coordinates n. When pressure, a change of resistance, or light is sensed at some image coordinate n_0, then the image is assigned the value 1:

$$g(n_0) = 1 \tag{10-5}$$

This continues until the user completes the drawing. These simple devices are quite useful for entering engineering drawings, handprinted characters, or other binary graphics in a binary image format.

Image Thresholding

Usually, a binary image is obtained from a gray-level image by some process of information abstraction. The advantage of the B-fold reduction in the required image storage space is offset by what can be a significant loss of information in the resulting binary image. However, if the process is accomplished with care, then a simple abstraction of information can be obtained that can enhance subsequent processing, analysis, or interpretation of the image.

The simplest such abstraction is the process of image thresholding, which can be thought of as an extreme form of gray-level quantization. Suppose that a gray-level image f can take K possible gray levels 0, 1, 2, ⋯, $K-1$. Define an integer threshold, T, that lies in the gray-scale range of $T E$ (0, 1, 2, ⋯, K−1).

The process of thresholding is a process of simple comparison: each pixel value in f is compared to T. Based on this comparison, a binary decision is made that defines the value of the corresponding pixel in an output binary image g:

$$g(n) = \begin{cases} 0 & if \geqslant f(n) \geqslant T \\ 1 & if ? f(n) < T \end{cases} \tag{10-6}$$

Of course, the threshold T that is used is of critical importance, since it controls the particular abstraction of information that is obtained. Indeed, different thresholds can produce different valuable abstractions of the image. Other thresholds may produce little valuable information at all. It is instructive to observe the result of thresholding an image at many different levels in sequence.

As will be seen, image thresholding can often produce a binary image result that is quite useful for simplified processing, interpretation, or display. However, some gray-level images do not lead to any interesting binary result regardless of the chosen threshold T.

Several questions arise: Given a gray-level image, how does one decide whether binari-



zation of the image by gray-level thresholding will produce a useful result? Can this be decided automatically by a computer algorithm? Assuming that thresholding is likely to be successful, how does one decide on a threshold level T? These are apparently simple questions pertaining to a very simple operation. However, these questions turn out to be quite difficult to answer in the general case. In other cases, the answer is simpler. In all cases, however, the basic tool for understanding the process of image thresholding is the image histogram.

Thresholding is most commonly and effectively applied to images that can be characterized as having bimodal histograms.

Figure 10.1 depicts two hypothetical image histograms. The one on the left has two clear modes; the one at the right either has a single mode, or two heavily overlapping, poorly separated modes. Bimodal histograms are often (but not always) associated with images that contain objects and backgrounds having a significantly different average brightness. This may imply bright objects on a dark background, or dark objects on a bright background.

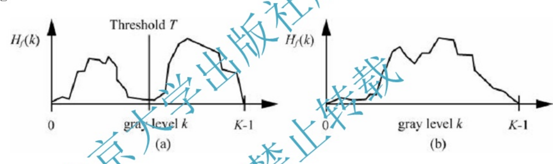


Figure 10.1 Hypothetical Histograms: (a) well-separated modes and (b) poorly separated or indistinct modes.

The goal, in many applications, is to separate the objects from the background, and to label them as object or as background. If the image histogram contains well-separated modes associated with an object and with a background, then thresholding can be the means for achieving this separation. Practical examples of gray-level images with well-separated bimodal histograms are not hard to find. For example, an image of machine-printed type (like that being currently read), or of handprinted characters, will have a very distinctive separation between object and background. Examples abound in biomedical applications, where it is often possible to control the lighting of objects and background. Standard bright-field microscope images of single or multiple cells (micrographs) typically contain bright objects against a darker background. In many industry applications, it is also possible to control the relative brightness of objects of interest and the backgrounds they are set against. For example, machine parts that are being imaged (perhaps in an automated inspection application) may be placed on a mechanical conveyor that has substantially

different reflectance properties than the objects.

New Words and Phrases

binary digital image 二值数字图像
Boolean *adj*. 布尔数学的
interpret *vt*. 解释，理解，诠释
equivalently *adv*. 等价地
coordinate *vt*. & *vi*. 使协调，使调和；整合
pressure sensitive tablet 压敏片
initialize *vt*. 初始化
acquisition *n*. 获得；取得
thresholding *n*. 阈值
binarization *n*. 二值化
pertain *vi*. 关于，有关；适合；附属
bimodal *adj*. 双峰的
depict *vt*. 描绘，描述
overlapping *n*. 重叠，搭接
label *n*. 标签
vt. 贴标签于；把……列为
conveyor *n*. 传送者；传送带
reflectance *n*. 反射比，反射系数

Exercises

Ⅰ. Answer the following questions.

1. What are demonstrate fundamental tools for the processing of binary digital images?
2. How are binary images displayed?
3. Where is a binary image obtained from?

Ⅱ. Translate the following sentences into Chinese.

1. Binary image processing is of special interest, since an image in binary formatcan be processed with very fast logical (Boolean) operators. Often, a binary image has been obtained by abstracting essential information from a gray-level image, such as object location, object boundaries, or the presence or absence of some image property.

____

____

____

____

____

2. Since most image display systems and software assume images of eight or more bits per pixel, the question arises as to how binary images are displayed. Usually, they are dis-



played using the two extreme gray tones, black and white, which are ordinarily represented by 0 and 255, respectively, in a gray-scale display environment.

____
____
____
____
____

3. Usually, a binary image is obtained from a gray-level image by some process of information abstraction. The advantage of the B-fold reduction in the required image storage space is offset by what can be a significant loss of information in the resulting binary image. However, if the process is accomplished with care, then a simple abstraction of information can be obtained that can enhance subsequent processing, analysis, or interpretation of the image.

____
____
____
____
____

Abstract Reading

Full-reference Quality Estimation for Images with Different Spatial Resolutions

Multimedia communication is becoming pervasive because of the progress in wireless communications and multimedia coding. Estimating the quality of the visual content accurately is crucial in providing satisfactory service. State of the art visual quality assessment approaches are effective when the input image and reference image have the same resolution. However, finding the quality of an image that has spatial resolution different than that of the reference image is still a challenging problem. To solve this problem, we develop a quality estimator (QE), which computes the quality of the input image without resampling the reference or the input images. In this paper, we begin by identifying the potential weaknesses of previous approaches used to estimate the quality of experience. Next, we design a QE to estimate the quality of a distorted image with a lower resolution compared with the reference image. We also propose a subjective test environment to explore the suc-

cess of the proposed algorithm in comparison with other QEs. When the input and test images have different resolutions, the subjective tests demonstrate that in most cases the proposed method works better than other approaches. In addition, the proposed algorithm also performs well when the reference image and the test image have the same resolution.

Segmentation-driven Image Registration-application to 4D DCE-MRI Recordings of the Moving Kidneys

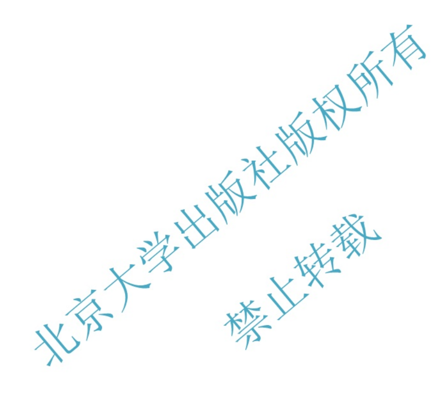
Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) of the kidneys requires proper motion correction and segmentation to enable an estimation of glomerular filtration rate through pharmacokinetic modeling. Traditionally, co-registration, segmentation, and pharmacokinetic modeling have been applied sequentially as separate processing steps. In this paper, a combined 4D model for simultaneous registration and segmentation of the whole kidney is presented. To demonstrate the model in numerical experiments, we used normalized gradients as data term in the registration and a Mahalanobis distance from the time courses of the segmented regions to a training set for supervised segmentation. By applying this framework to an input consisting of 4D image time series, we conduct simultaneous motion correction and two-region segmentation into kidney and background. The potential of the new approach is demonstrated on real DCE-MRI data from ten healthy volunteers.

Model-based Edge Detector for Spectral Imagery Using Sparse Spatiospectral Masks

Two model-based algorithms for edge detection in spectral imagery are developed that specifically target capturing intrinsic features such as is oluminant edges that are characterized by a jump in color but not in intensity. Given prior knowledge of the classes of reflectance or emittance spectra associated with candidate objects in a scene, a small set of spectral-band ratios, which most profoundly identify the edge between each pair of materials, are selected to define a edge signature. The bands that form the edge signature are fed into a spatial mask, producing a sparse joint spatiospectral nonlinear operator. The first algorithm achieves edge detection for every material pair by matching the response of the operator at every pixel with the edge signature for the pair of materials. The second algorithm is a classifier-enhanced extension of the first algorithm that adaptively accentuates distinctive features before applying the spatiospectral operator. Both algorithms are extensively verified using spectral imagery from the airborne hyperspectral imager and from a dots-in-a-



well midinfrared imager. In both cases, the multicolor gradient (MCG) and the hyperspectral/spatial detection of edges (HySPADE) edge detectors are used as a benchmark for comparison. The results demonstrate that the proposed algorithms outperform the MCG and HySPADE edge detectors in accuracy, especially when isoluminant edges are present. By requiring only a few bands as input to the spatiospectral operator, the algorithms enable significant levels of data compression in band selection. In the presented examples, the required operations per pixel are reduced by a factor of 71 with respect to those required by the MCG edge detector.



UNIT 11

VIDEO PROCESSING

Text: Video Segmentation

Video segmentation refers to the identification of regions in a frame of video that are homogeneous in some sense. Different features and homogeneity criteria generally lead to different segmentations of the same data; for example, color segmentation, texture segmentation, and motion segmentation usually result in different segmentation maps. Furthermore, there is no guarantee that any of the resulting segmentations will be semantically meaningful, since a semantically meaningful region may have multiple colors, multiple textures, or multiple motion.

In this paper, we are primarily concerned with labeling independently moving image regions (motion segmentation) or semantically meaningful image regions (video object plane segmentation). Motion segmentation (also known as optical flow segmentation) methods label pixels (or optical flow vectors) at each frame that are associated with independently moving part of a scene. These regions may or may not be semantically meaningful. For example, a single object with articulated motion may be segmented into multiple regions. Although it is possible to achieve fully automatic motion segmentation with some limited accuracy, semantically meaningful video object segmentation generally requires user to define the object of interest in at least some key frames. Motion segmentation is closely related to two other problems, motion (change) detection and motion estimation. Change detection is a special case of motion segmentation with only two regions, namely changed and unchanged regions (in the case of a static camera) or global and local motion regions (in the case of a moving camera). An important distinction between change detection and motion segmentation is that the former can be achieved without motion estimation if the scene is recorded with a static camera. Change detection in the case of a moving camera and



general motion segmentation, in contrast, require some sort of global or local motion estimation, either explicitly or implicitly. Motion detection and segmentation are also plagued with the same two fundamental limitations associated with motion estimation: occlusion and aperture problems.

For example, pixels in a flat image region may appear stationary even if they are moving as a result of an aperture problem (hence the need for hierarchical methods); or erroneous labels may be assigned to pixels in covered or uncovered image regions as a result of an occlusion problem.

It should not come as a surprise that motion/object segmentation is an integral part of many video analysis problems, including (i) improved motion (optical flow) estimation, (ii) three-dimensional (3-D) motion and structure estimation in the presence of multiple moving objects, and (iii) description of the temporal variations or the content of video. In the former case, the segmentation labels help to identify optical flow boundaries (motion edges) and occlusion regions where the smoothness constraint should be turned off. Segmentation is required in the second case, because distinct 3-D motion and structure parameters are needed to model the flow vectors associated with each independently moving object. Finally, in the third case, segmentation information may be employed in an object-level description of frame-to-frame motion, as opposed to a pixel-level description provided by individual flow vectors.

As with any segmentation problem, proper feature selection facilitates effective motion segmentation. In general, the application of standard image segmentation methods directly to estimated optical flow vectors may not yield meaningful results, since an object moving in 3-D usually generates a spatially varying optical flow field. For example, in the case of a rotating object, there is no flow at the center of the rotation, and the magnitude of the flow vectors grows as we move away from the center of rotation. Therefore, a parametric model-based approach, where we assume that the motion field can be described by a set of K parametric models, is usually adopted. In parametric motion segmentation, the model parameters are the motion features.

Then, motion segmentation algorithmsaim to determine the number of motion models that can adequately describe a scene, type/complexity of these motion models, and the spatial support of each motion model. The most commonly used types of parametric models are affine, perspective, and quadratic mappings, which assume a 3-D planar surface in motion. In the case of a nonplanar object, the resulting optical flow can be modeled by a piecewise affme, perspective, or quadratic flow field if we approximate the object surface by a union of a small number of planar patches. Because each independently moving object or planar patch will best fit a different parametric model, the parametric approach may lead

to an oversegmentation of motion in the case of nonplanar objects.

It is difficult to associate a generic figure of merit with a video segmentation result. If segmentation is employed to improve the compression efficiency or rate control, then oversegmentation may not be a cause of concern. The occlusion and aperture problems are mainly responsible for misalignment of motion and actual object boundaries. Furthermore, model misfit possibly asa result of a deviation of the surface structure from a plane generally leads to oversegmentation of the motion field.

In contrast, if segmentation is needed for object-based editing and composition as in the upcoming MPEG-4[1] standard, then it is of utmost importance that the estimated boundaries align with actual object boundaries perfectly. Even a single pixel error may not be tolerable in this case. Although elimination of outlier motion estimates and imposing spatio temporal smoothness constraints on the segmentation map improve the chances of obtaining more meaningful segmentation results, semantic object segmentation in general requires specialized capture methods (chroma keying) or user interaction (semi-automatic methods).

New Words and Phrases

segmentation *n.* 分割；分段
frame *n.* 框架，组织
homogeneous *adj.* 同性质的，同类的
criteria *n.* 标准，准则(criterion 的复数)
semantically *adv.* 语义地
optical *adj.* 视觉的，光学的
vector *n.* 向量
articulate *adj.* 发音清晰的；善于表达的；有关节的
vt. 清晰地发(音)；言语表达；(用关节)连接
explicitly *adv.* 明白地，明确地
implicitly *adv.* 含蓄地，暗示地
plague *vt.* 使痛苦，造成麻烦
occlusion *n.* 闭塞；咬合；堵塞
aperture *n.* 孔，洞；光圈
stationary *adj.* 静止的，平衡的
hierarchical *adj.* 分层的
erroneous *adj.* 错误的
temporal *adj.* 时间的；暂存的
parameter *n.* 参数
facilitate *vt.* 促进，助长
magnitude *n.* 巨大，重大；量级
parametric *adj.* 参数的，参量的
affine *adj.* 仿射的
quadratic *adj.* 二次的
n. 二次方程式
nonplanar *adj.* 非平面的，空间的
planar *adj.* 平面的
oversegmentation *n.* 重复分割
merit *n.* 价值，优点
vt. 值得，应获得
misalignment *n.* 未对准，不重合；位移
misfit *vt.* 对……不适合；不适用于
n. 不适合
deviation *n.* 背离，偏离，偏差



chroma *n*. 色度，饱和度

Notes

1. MPEG-4：MPEG 于 1999 年 2 月正式公布了 MPEG-4(ISO/IEC 14496)标准第一版本。同年年底 MPEG-4 第二版问世，且于 2000 年年初正式成为国际标准。MPEG-4 与 MPEG-1 和 MPEG-2 有很大的不同。MPEG-4 不只是具体的压缩算法，它是针对数字电视、交互式绘图应用(影音合成内容)、交互式多媒体(WWW、资料撷取与分散)等整合及压缩技术的需求而制定的国际标准。MPEG-4 标准将众多多媒体应用集成于一个完整框架内，旨在为多媒体通信及应用环境提供标准算法及工具，从而建立起一种能被多媒体传输、存储、检索等应用领域普遍采用的统一数据格式。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. motion segmentation
2. video object plane segmentation
3. optical flow vector
4. static camera
5. frame-to-frame motion
6. planar patch
7. chroma keying
8. semi-automatic method

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

semantically	representation	semantic	classified
features	automatic	clustering	intensity
samples	occurrence	technology	values
alternative	flow	definition	texture

1. The Hough transform is a well-known ________ technique in which the data ________ "vote" for the most representative feature ________ in a quantized feature space.

2. Motion segmentation approaches in general are ________ as optical ________ segmentation methods, which operate on precomputed optical flow estimates as an input

________, and direct methods, which operate on spatiotemporal ________ values.

3. It is difficult to achieve ________ meaningful object segmentation by using fully ________ methods based on low-level features such as motion, color, and ________. This is because a ________ object may contain multiple motions, colors, textures, and so on, and the ________ of semantic objects may depend on the context, which may not be possible to capture by using low-level ________.

4. Chroma keying is an object-based video ________ in which each video object is captured individually in a special studio against a key color.

5. Because chroma keying requires special studios or equipment to capture video objects, an ________ approach is interactive segmentation, using automated tools to aid a human operator. To this effect, we assume that the contour of the first ________ of the semantic object of interest is marked interactively by a human operator.

Ⅲ. Translate the following sentences into Chinese.

1. Segmentation by dominant motion analysis refers to extracting one object (with the dominant motion) from the scene at a time. Dominant motion segmentation can be considered as a hierarchically structured top-down approach, which starts by fitting a single parametric motion model to the entire frame, and then partitions the frame into two regions, those pixels that are well represented by this dominant motion model and those that are not.

____
____
____
____
____

2. Multiple object segmentation can be achieved by repeating the same procedure on the residual image after each object is extracted. Once the first dominant object is segmented and tracked, the procedure can be repeated recursively to segment and track the next dominant object after excluding all pixels belonging to the first object from the region of analysis. Hence, the method is capable of segmenting multiple moving objects in a top-down fashion if a dominant motion exists at each stage.

____
____
____
____
____



3. Multiple motion segmentation methods let multiple motion models compete against each other at each decision site. They consist of three basic steps, which are strongly interrelated: estimation of the number K of independent motions, estimation of model parameters for each motion, and determination of support of each model (segmentation labels).

Reading: Video Compression Application Requirements

Awide variety of digital video applications currently exist. They range from simple low-resolution and low-bandwidth applications (multimedia, picturephone) to very high-resolution and high-bandwidth (HDTV) demands. This paper will present requirements of current and future digital video applications and the demands they place on the video compression system.

Asa way to demonstrate the importance of video compression, the transmission of digital video television signals is presented. The bandwidth required by a digital television signal is approximately one-half the number of picture elements (pixels) displayed per second. The analog pixel size in the vertical dimension is the distance between scanning lines, and the horizontal dimension is the distance the scanning spot moves during 1/2 cycle of the highest video signal transmission frequency. The bandwidth is given by:

$$B_W = 0.8(F_R)(M_L)(R_H) \quad (11\text{-}1)$$

where

B_W = system bandwidth,

F_R = number of frames transmitted per second (fps),

N_L = number of scanning lines per frame,

R_H = horizontal resolution (lines), proportional to pixel resolution.

The National Television Systems Committee (NTSC) aspect ratio is 4/3, the constant 0.5 is the ratio of the number of cycles to the number of lines, and the factor 0.84 is the fraction of the horizontal scanning interval that is devoted to signal transmission.

The NTSC transmission standard used for television broadcasts in the United States has the following parameter values: F_R=29. 97 fps, N_L=525 lines, and R_H=340 lines. This yields a video system bandwidth B_W of 4.2 MHz for the NTSC broadcast system. In

order to transmit a color digital video signal, the digital pixel format must be defined. The digital color pixel is made of three components: one luminance (Y) component occupying 8 bits, and two color difference components (U and V) each requiring 8 bits. The NTSC picture frame has 720×480×2 total luminance and color pixels. In order to transmit this information for an NTSC broadcast system at 29.97 frames/s, the following bandwidth is required

$$\text{Digital } B_W \approx \frac{\text{bitrate}}{2}$$

$$= \frac{(29.97\ \text{fps}) \times (24\ \text{bits/pixel}) \times (720 \times 480 \times 2\ \text{pixels/frame})}{2}$$

$$= 249\text{MHz} \tag{11-2}$$

This represents an increase of ~59 times the available system bandwidth, and ~41 times the full transmission channel bandwidth (6 MHz) for current NTSC signals. HDTV picture resolution requires up to three times more raw bandwidth than this example! (Two transmission channels totaling 12 MHz are allocated for terrestrial HDTV transmissions.) It is clear from this example that terrestrial television broadcast systems will have to use digital transmission and digital video compression to achieve the overall bitrate reduction and image quality required for HDTV signals.

The example not only points out the significant bandwidth requirements for digital video information, but also indirectly brings up the issue of digital video quality requirements. The tradeoff between bitrate and quality or distortion is a fundamental issue facing the design of video compression systems.

To this end, it is important to fully characterize an application's video communications requirements before designing or selecting an appropriate video compression system. Factors that should be considered in the design and selection of a video compression system include the following items.

1. Video characteristics: video parameters such as the dynamic range, source statistics, pixel resolution, and noise content can affect the performance of the compression system.

2. Transmission requirements: transmission bitrate requirements determine the power of the compression system. Very high transmission bandwidth, storage capacity, or quality requirements may necessitate lossless compression. Conversely, extremely low bitrate requirements may dictate compression systems that trade off image quality for a large compression ratio. Progressive transmission is a key issue for selection of the compression system. It is generally used when the transmission bandwidth exceeds the compressed video bandwidth. Progressive coding refers to a multiresolution, hierarchical, or subband en-



coding of the video information. It allows for transmission and reconstruction of each resolution independently from low to high resolution. In addition, channel errors affect system performance and the quality of the reconstructed video. Channel errors can affect the bit stream randomly or in burst fashion. The channel error characteristics can have different effects on different encoders, and they can range from local to global anomalies. In general, transmission error correction codes (ECC) are used to mitigate the effect of channel errors, but awareness and knowledge of this issue is important.

3. Compression system characteristics and performance: the nature of video applications makes many demands on the video compression system. Interactive video applications such as videoconferencing demand that the video compression systems have symmetric capabilities. That is, each participant in the interactive video session must have the same video encoding and decoding capabilities, and the system performance requirements must be met by both the encoder and decoder. In contrast, television broadcast video has significantly greater performance requirements at the transmitter because it has the responsibility of providing real-time high quality compressed video that meets the transmission channel capacity. Digital video system implementation requirements can vary significantly. Desktop televideo conferencing can be implemented by using software encoding and decoding, or it may require specialized hardware and transmission capabilities to provide a high-quality performance. The characteristics of the application will dictate the suitability of the video compression algorithm for particular system implementations. The importance of the encoder and system implementation decision cannot be overstated; system architectures and performance capabilities are changing at a rapid pace and the choice of the best solution requires careful analysis of the all possible system and encoder alternatives.

4. Rate-distortion requirements: the rate-distortion requirement is a basic consideration in the selection of the video encoder. The video encoder must be able to provide the bitrate(s) and video fidelity (or range of video fidelity) required by the application. Otherwise, any aspect of the system may not meet specifications. For example, if the bitrate specification is exceeded in order to support a lower MSE, a larger than expected transmission error rate may cause a catastrophic system failure.

5. Standards requirements: video encoder compatibility with existing and future standards is an important consideration if the digital video system is required to interoperate with existing or future systems. A good example is that of a desktop videoconferencing application supporting a number of legacy video compression standards. This results in requiring support of the older video encoding standards on new equipment designed for a newer incompatible standard. Videoconferencing equipment not supporting the old standards would not be capable or as capable to work in environments supporting older standards.

New Words and Phrases

vertical *adj*. 垂直的，直立的
　　n. 垂直线，垂直面
horizontal *adj*. 水平的；地平线的；同一阶层的
　　n. 水平线，水平面；水平位置
fraction *n*. 分数；部分
terrestrial *adj*. 地球的；陆地的
　　n. 陆地生物；地球上的人
bitrate *n*. 比特率；位率；位速率
necessitate *vt*. 使成为必需，需要；迫使
multiresolution *n*. 多分辨率
subband *n*. 部分波段；次能带
anomaly *n*. 异常；不规则；反常事物
ECC *abbr*. 纠错码(Error Correction Code)
symmetric *adj*. 对称的；匀称的
rate-distortion *n*. 率失真
fidelity *n*. 保真度
catastrophic *adj*. 灾难的；悲惨的；灾难性的，毁灭性的

Exercises

Ⅰ. Answer the following questions.

1. What factors should be considered in the design and selection of a video compression system?
2. What is video characteristic?
3. What is the NTSCT transmission standard?

Ⅱ. Translate the following sentences into Chinese.

1. A wide variety of digital video applications currently exist. They range from simple low-resolution and low-bandwidth applications (multimedia, Picturephone) to very high-resolution and high-bandwidth (HDTV) demands.

2. As a way to demonstrate the importance of video compression, the transmission of digital video television signals is presented. The bandwidth required by a digital television signal is approximately one-half the number of picture elements (pixels) displayed per second.



3. The nature of video applications makes many demands on the video compression system. Interactive video applications such as videoconferencing demand that the video compression systems have symmetric capabilities. That is, each participant in the interactive video session must have the same video encoding and decoding capabilities, and the system performance requirements must be met by both the encoder and decoder.

Abstract Reading

Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines

We present a model for building, visualizing, and interacting with multiscale representations of information visualization techniques using hierarchical aggregation. The motivation for this work is to make visual representations more visually scalable and less cluttered. The model allows for augmenting existing techniques with multiscale functionality, as well as for designing new visualization and interaction techniques that conform to this new class of visual representations. We give some examples of how to use the model for standard information visualization techniques such as scatterplots, parallel coordinates, and node-link diagrams, and discuss existing techniques that are based on hierarchical aggregation. This yields a set of design guidelines for aggregated visualizations. We also present a basic vocabulary of interaction techniques suitable for navigating these multiscale visualizations.

Improving the Efficiency of Viewpoint Composition

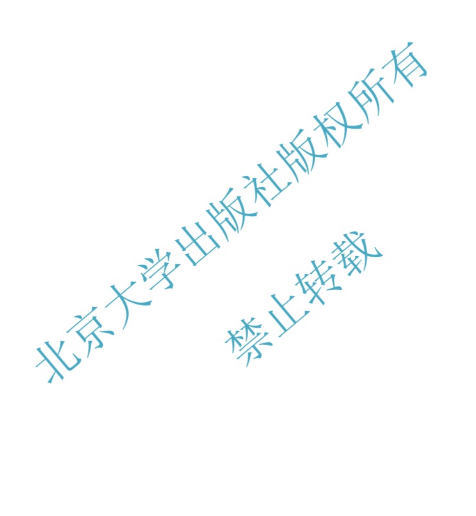
In this paper, we concentrate on the problem of finding the viewpoint that best satisfies a set of visual composition properties, often referred to as Virtual Camera or Viewpoint Composition. Previous approaches in the literature, which are based on general optimization solvers, are limited in their practical applicability because of unsuitable computation times and limited experimental analysis. To bring performances much closer to the needs of interactive applications, we introduce novel ways to define visual properties, evaluate their satisfaction, and initialize the search for optimal viewpoints, and test them in several problems under various time budgets, quantifying also, for the first time in the domain, the importance of tuning the parameters that control the behavior of the solving process. While our solver, as others in the literature, is based on Particle Swarm Optimization, our contributions could be applied to any stochastic search process that solves through many viewpoint evaluations, such as the genetic algorithms employed by other papers in the literature. The complete source code of our approach, together with the scenes and problems we have employed, can be downloaded from https://bitbucket.org/rranon/smart-viewpoint-computation-lib.

Assessing the Ability of a VR-based Assembly Task Simulation to Evaluate Physical Risk Factors

Nowadays, the process of workstation design tends to include assessment steps in a virtual environment (VE) to evaluate the ergonomic features. These approaches are cost-effective and convenient since working directly on the digital mock-up in a VE is preferable to constructing a real physical mock-up in a real environment (RE). This study aimed at understanding the ability of a VR-based assembly tasks simulator to evaluate physical risk factors in ergonomics. Sixteen subjects performed simplified assembly tasks in RE and VE. Motion of the upper body and five muscle electromyographic activities were recorded to compute normalized and averaged objective indicators of discomfort, that is, rapid upper limb assessment score, averaged muscle activations, and total task time. Rated perceived exertion (RPE) and a questionnaire were used as subjective indicators of discomfort. The timing regime and complexity of the assembly tasks were investigated as within-subject factors. The results revealed significant differences between measured indicators in RE and VE. While objective measures indicated lower activity and exposure in VE, the subjects experienced more discomfort than in RE. Fairly good correlation levels were found between



RE and VE for six of the objective indicators. This study clearly demonstrates that ergonomic studies of assembly tasks using VR are still challenging. Indeed, objective and subjective measurements of discomfort that are usually used in ergonomics to minimize the risks of work-related musculoskeletal disorders development exhibit opposite trends in RE and VE. Nevertheless, the high level of correlation found during this study indicates that the VR-based simulator can be used for such assessments.



UNIT 12

WIRELESS COMMUNICATION

Text: Speech Coding

Introduction

1. Speech Telephony as Conversational Multimedia Service

When O. Nußbaumer succeeded in the first wireless transmission of speech and music in the experimental physics lab at Graz University of Technology in 1904, nobody would have predicted the tremendous growth in wireless multimedia communications 100 years after this historical achievement.

Many new media types have emerged such as text, image, and video, and modern services range from essentially one-way media download, browsing, messaging, application sharing, broadcasting, and real-time streaming to two-way, interactive, real-time conversations by text (chat), speech, and video telephony. Still, speech telephony is the backbone of all conversational service and continues as an indispensable functionality of any mobile communications terminal. It is for this reason that we will focus our discussion of source-coding methods on speech signals—i. e., on their efficient digital representation for transmission (or storage) applications.

The success story of digital speech coding started with the introduction of digital switching in the Public Switched Telephone Network (PSTN)[1] using Pulse Code Modulated (PCM)[2] speech at 64 Kbps and continued with a cascade of advanced compression standards from 32 Kbps in the early 1980s over 16 Kbps to 8 Kbps in the late 1990s, all with a focus on long-distance circuit multiplication while maintaining the traditional high-

quality level of wireline telephony (toll quality). For wireless telephony, the requirements on digital coding were rather stringent with regard to bit rate and complexity from the very beginning in the mid-1980s whereas some compromises in speech quality seemed acceptable because users either had never made the experience of mobile telephony before or, if so, their expectations were biased from the relatively poor quality of analog mobile radio systems. This situation changed quickly, and the first standards introduced at the beginning of the 1990s were completely overturned within 5 years with significant quality enhancements through new coding algorithms, maintaining a bit rate of about 12 Kbps where the accompanying complexity increase was mitigated by related advances in microelectronics and algorithm implementation.

2. Source-coding Basics

The foundations for source coding were laid by C. Shannon [1959], who developed not only channel-coding theory for imperfect transmission channels but also rate-distortion theory for signal compression. The latter theory is based on two components.

(1) a stochastic source model which allows us to characterize the redundancy in source information.

(2) a distortion measure which characterizes the relevance of source information for a user. For asymptotically infinite delay and complexity, and certain simple source models and distortion measures, it can be shown that there exists an achievable lower bound on the bit rate necessary to achieve a given distortion level and, vice versa, that their exists an achievable lower bound on the distortion to be tolerated for a given bit rate. While complexity is an ever-dwindling obstacle, delay is a substantial issue in telephony, as it degrades the interactive quality of conversations severely when it exceeds a few 100 ms. Therefore, the main insight from rate-distortion theory is the existence of a three-way tradeoff among the fundamental parameters rate, distortion, and delay. Traditional telephony networks operate in circuit-switched mode where transmission delay is essentially given by the electromagnetic propagation time and becomes only noticeable when dealing with satellite links. However, packet-switched networks are increasingly being used for telephony as well—as in Voice over Internet Protocol (VoIP) systems—where substantial delays can be accumulated in router queues, etc.. In such systems, delay becomes the most essential parameter and will determine the achievable rate-distortion tradeoff.

Source coding with a small but tolerable level of distortion is also known as lossy coding whereas the limiting case of zero distortion is known as lossless coding. In most cases, a finite rate allows lossless coding only for discrete amplitude signals which we might consider for transcoding of PCM speech—i. e. , the digital compression of speech signals which

have already been digitized with a conventional PCM codec. However, for circuit-switched wireless speech telephony, such lossless coders have two drawbacks: first, they waste the most precious resource—i. e. , the allocated radio spectrum—as they invest more bits than necessary to meet the quality expectations of a typical user; second, they often result in a bit stream with a variable rate—e. g. , when using a Huffman coder—which cannot be matched efficiently to the fixed rate offered by circuit switched transmission.

Variable-rate coding is, however, a highly relevant topic in packet-switched networks and in certain applications of joint source channel coding for circuit-switched networks. While Shannon's theory shows that, under idealized conditions, source coding and channel coding can be fully separated such that the two coding steps can be designed and optimized independently, this is not true under practical constraints such as finite delay or time-varying conditions where only a joint design of the source and channel coders is optimal. In this case, a fixed rate offered by the network can be advantageously split into a variable source rate and a variable channel code rate.

3. Speech Coder Designs

Source-coding theory teaches us how to use models of source redundancy and of user-defined relevance in the design of speech-coding systems. Perceptual relevance aspects will be discussed later; however, the use of the source model gives rise to a generic classification of speech coder designs.

(1) Waveform coders use source models only implicitly to design an adaptive dynamical system which maps the original speech waveform on a processed waveform that can be transmitted with fewer bits over the given digital channel. The decoder essentially inverts encoder processing to restore a faithful approximation of the original waveform. All waveform coders share the property that an increase of the bit rate will asymptotically result in lossless transcoding of the original PCM waveform. For such systems, the definition of a coding error signal as the difference between the original and the decoded waveform makes sense (although it is no immediate measure of the perceptual relevance of the distortion introduced).

(2) Model-based coders or vocoders rely on an explicit source model to represent the speech signal using a small set of parameters which the encoder estimates, quantizes, and transmits over the digital channel. The decoder uses the received parameters to control a real-time implementation of the source model that generates the decoded speech signal. An increase of the bit rate will result in saturation of the speech quality at a nonzero distortion level which is limited by systematic errors in the source model. Only recently, model-based coders have advanced to a level where these errors have little perceptual impact, allowing their use for very-low-rate applications (2. 4 Kbps and below) with slightly reduced quality



constraints. Furthermore, due to the signal generation process in the decoder, the decoded waveform is not synchronized with the original waveform and, therefore, the definition of a waveform error is useless to characterize the distortion of model-based coders.

(3) Hybrid coders aim at the optimal mix of the two previous designs. They start out with a model based approach to extract speech signal parameters but still compute the modeling error explicitly on the waveform level. This model error or residual waveform is transmitted using a waveform coder whereas the model parameters are quantized and transmitted as side information. The two information streams are combined in the decoder to reconstruct a faithful approximation of the waveform such that hybrid coders share the asymptotically lossless coding property with waveform coders. Their advantage lies in the explicit parameterization of the speech model which allows us to exploit more advanced models than is the case with pure waveform coders which rely on a single invertible dynamical system for their design.

The model-based view of speech-coding design suggests that description of a speech-coding system should always start with the decoder that typically contains an implementation of the underlying speech model. The encoder is then obtained as the signal analysis system that extracts the relevant model parameters and residual waveform. Therefore, the encoder has a higher complexity than the decoder and is more difficult to understand and implement. In this sense, a speech-coding standard might specify only the decoder and the format for transmitted data streams while leaving the design of the best matching encoder to industrial competition.

Further Design Issues

The reliance on source models for speech coder design naturally results in a dependence of coder performance on the match between this model and the signal to be encoded. Any signal that is not clean speech produced from a single talker near the microphone may suffer from additional distortion which requires additional performance testing and, possibly, design modifications. Examples of these extra issues are the suitability of a coder for music (e. g. , if put on hold while waiting for a specific party), for severe acoustic background noise (e. g. , talking from the car, maybe with open windows), for babble noise from other speakers (e. g. , talking from a cafeteria), for reverberation (e. g. , talking in hands-free mode), etc. .

Further system design aspects include the choice between narrowband speech as used in traditional wireline telephony (e. g. , an analog bandwidth from 300 Hz to 3. 4 kHz with 8-kHz sampling frequency) and wideband speech with quality similar to Frequency Modula-

tion (FM) radio (e. g. , an analog bandwidth from 50Hz to 7 KHz with 16-KHz sampling frequency), which substantially enhances user experience with clearly noticeable speech quality improvements over and above the Plain Old Telephone Service (POTS). Second, even with the use of sophisticated channel codes, the robustness of channel errors such as individual bit errors, bursts, or entire lost transmission frames or packets is also a source-coding design issue.

Finally, within the network path from the talker to the listener, there may be several codec-tandeming steps where transcoding from one coding standard to another occurs, each time with a potential further loss of speech quality.

New Words and Phrases

browse *vt*. & *vi*. 浏览；随意翻阅
n. 浏览
telephony *n*. 拨号服务；电话语音
backbone *n*. 分水岭；支柱
indispensable *adj*. 不可缺少的
n. 不可缺少的人或物
functionality *n*. (计算机或电子系统的)功能；功能性
cascade *n*. 级联；串联
toll *vt*. 向……征收捐税
vi. 鸣钟；收费
n. 税，通行税；通行费
stringent *adj*. 严格的；迫切的
compromise *n*. 妥协
vi. 折中解决；妥协
vt. 违背(原则)
bias *n*. 偏差；背离率指标
analog *n*. 模拟；相似体
adj. 模拟的
overtune *n*. 序曲
enhancement *n*. 增益
asymptotically *adv*. 渐近地
dwindling *adj*. 逐渐减少的
trade off 权衡
propagation *n*. 传播，传输，蔓延
codec *n*. 编码译码器
allocate *vt*. 分配，分派；把……拨给
spectrum *n*. 谱，光谱
split *vt*. 分裂；分开
n. 划分，分歧
adj. 裂开的
perceptual *adj*. 知觉的，有知觉的；感性
invert *vt*. 使……前后倒置；使反转
approximation *n*. 近似值；近似
vocoder *n*. 声码器；自动语音合成仪
waveform *n*. 波形
hybrid *n*. 杂种；混合物；混合词
adj. 混合的；杂种的
residual *adj*. 残余的；残留的
n. 剩余；残渣
modification *n*. 修改，修正，变更
acoustic *adj*. 听觉的；声学的
babble *vi*. 喋喋不休
vt. 含糊不清地说；泄露
n. 胡言乱语
reverberation *n*. 混响；残响；回响；反射
hands-free mode 免提模式
narrowband *n*. 窄带



modulation *n.* 调制；调幅度

Notes

1. 公用电话交换网(PSTN，Public Switch Telephone Network)，即日常生活中常用的电话网。众所周知，PSTN 是一种以模拟技术为基础的电路交换网络。在众多的广域网互联技术中，通过 PSTN 进行互连所要求的通信费用最低，但其数据传输质量及传输速度也最差，同时 PSTN 的网络资源利用率也比较低。

2. 脉冲编码调制(Pulse Code Modulation)，简称 PCM。是对连续变化的模拟信号进行抽样、量化和编码产生的数字信号。PCM 的优点是音质好，缺点是体积大。PCM 可以提供用户从 2 Mbps 到 155 Mbps 速率的数字数据专线业务，也可以提供话音、图像传送、远程教学等其他业务。

Exercises

Ⅰ. Please translate the following words and phrases into Chinese.

1. source-coding method
2. analog mobile radio system
3. rate-distortion theory
4. ever-dwindling obstacle
5. packet-switched network
6. VoIP system
7. Variable-rate coding
8. circuit-switched network
9. hybrid coder
10. Frequency Modulation

Ⅱ. Fill in the blanks with the missing word(s) from the table below.

components	mechanical	encoded	dynamic
performance	vocal	converts	remarkable
audio	independent	relevant	frequency
distortion	representation	evaluation	collapsed
filtering	perception	imprecise	modifications



1. The ultimate recipient of human speech is the human hearing system, a ________ receiver with two broadband, directional antennas shaped for spatiotemporal ________ (the outer ear) in terms of the individual, monaural Head Related Transfer Functions, HRTFs (functions of both azimuth angle and ________) which along with interaural delay ________ give rise to our spatial hearing ability.

2. A highly adaptive, ________ impedance-matching network (the middle ear) covers a ________ range of more than 100 dB and a 3000-channel, phase-locking, threshold-based receiver (hair cells in the inner ear's cochlea) with very low self-noise (just above the level where we could hear our own blood-flow-induced noise) ________ the signal to an extremely parallel, low rate, synchronized ________ useful for distributed processing with low-power and ________ circuits (our nervous system).

3. Auditory models in the form of psychoacoustic—i.e., behavioral-models of ________ are very popular in ________ coding (like those for the ubiquitous MP3 standard) because they allow us to separate ________ from irrelevant parts of the information. For instance, certain signal ________ may be masked by others to such an extent that they become totally inaudible.

4. In particular, all the natural sound sources (which can be located at many positions along the vocal tract and which are often controlled by the local aerodynamic flow) are ________ into a single source which drives the filter in an ________ way. Furthermore, there is only a single output whereas the natural production system may switch between or even combine the oral and nasal branches of the ________ tract.

5. The reliance on source models for speech coder design naturally results in a dependence of coder ________ on the match between this model and the signal to be ________. Any signal that is not clean speech produced from a single talker near the microphone may suffer from additional ________ which requires additional performance testing and, possibly, design ________.

Ⅲ. Translate the following sentences into Chinese.

1. While the "sound of music" includes a wide range of signal generation mechanisms as provided by an orchestra of musical instruments, the instrument for generating speech is fairly unique and constitutes the physical basis for speech modeling, even at the acoustic or perception levels.

____

____

____

____



2. In a nutshell, speech communication consists of information exchange using a natural language as its code and the human voice as its carrier. Voice is generated by an intricate oscillator—the vocal folds—which is excited by sound pressure from the lungs.

3. The proof of a speech coder lies in listening. Till today, the best way of evaluating the quality of a speech coder is by controlled listening tests performed with sizable groups of listeners (a couple of dozens or more).

Reading: Applications and Requirements of Wireless Services

Wireless communications is one of the big engineering success stories of the last 25 years—not only from a scientific point of view, where the progress has been phenomenal but also in terms of market size and impact on society. Companies that were completely unknown 25 years ago are now household names all over the world, due to their wireless products, and in several countries the wireless industry is dominating the whole economy. Working habits, and even more generally the ways we all communicate, have been changed by the possibility of talking "anywhere, anytime".

For a long time, wireless communications has been associated with cellular telephony, as this is the biggest market segment, and has had the highest impact on everyday lives. In recent times, wireless computer networks have also led to a significant change in working habits and mobility of workers—answering emails in a coffee shop has become an everyday occurrence. But besides these widely publicized cases, a large number of less obvious applications have been developed, and are starting to change our lives. Wireless sensor networks monitor factories, wireless links replace the cables between computers and keyboards, and wireless positioning systems monitor the location of trucks that have goods i-

dentified by wireless Radio Frequency (RF) tags. This variety of new applications causes the technical challenges for the wireless engineers to become bigger with each day. This book aims to give an overview of the solution methods for current as well as future challenges.

Quite generally, there are two paths to developing new technical solutions: engineering driven and market driven. In the first case, the engineers come up with a brilliant scientific idea—without having an immediate application in mind. As time progresses, the market finds applications enabled by this idea. In the other approach, the market demands a specific product and the engineers try to develop a technical solution that fulfills this demand. In this paper, we describe these market demands. We start out with a brief history of wireless communications, in order to convey a feeling of how the science, as well as the market, has developed in the past 100 years. Then follows a description of the types of services that constitute the majority of the wireless market today. Each of these services makes specific demands in terms of data rate, range, number of users, energy consumption, mobility, and so on.

History

1. How It All Started

When looking at the history of communications, we find that wireless communications is actually the oldest form—shouts and jungle drums did not require any wires or cables to function. Even the oldest "electromagnetic" (optical) communications are wireless: smoke signals are based on propagation of optical signals along a line-of-sight connection. However, wireless communications as we know it started only with the work of Maxwell and Hertz, who laid the basis for our understanding of the transmission of electromagnetic waves. It was not long after their groundbreaking work that Tesla demonstrated the transmission of information via these waves—in essence, the first wireless communications system. In 1898, Marconi made his well-publicized demonstration of wireless communications from a boat to the Isle of Wight in the English Channel. It is noteworthy that while Tesla was the first to succeed in this important endeavor, Marconi had the better public relations, and is widely cited as the inventor of wireless communications, receiving a Nobel prize in 1909.

In the subsequent years, the use of radio (and later television) became widespread throughout the world. While in the "normal" language, we usually do not think of radio or TV as "wireless communications", they certainly are, in a scientific sense, information



transmission from one place to another by means of electromagnetic waves. They can even constitute "mobile communications", as evidenced by car radios. A lot of basic researches—especially concerning wireless propagation channels—were done for entertainment broadcasting. By the late 1930s, a wide network of wireless information transmission—though unidirectional—was in place.

2. The First Systems

At the same time, the need for bidirectional mobile communications emerged. Police departments and the military had obvious applications for such two-way communications, and were the first to use wireless systems with closed user groups. Military applications drove a lot of the research during, and shortly after, the Second World War. This was also the time when much of the theoretical foundations for communications in general were laid. Claude Shannon's groundbreaking work A Mathematical Theory of Communications appeared during that time, and established the possibility of error-free transmission under restrictions for the data rate and the Signal-to-Noise Ratio (SNR). Some of the suggestions in that work, like the use of optimum power allocation in frequency-selective channels, are only now being introduced into wireless systems.

The 1940s and 1950s saw several important developments: the use of Citizens' Band (CB) radios became widespread, establishing a new way of communicating between cars on the road. Communicating with these systems was useful for transferring vital traffic information and related aspects within the closed community of the drivers owning such devices, but it lacked an interface to the public telephone system, and the range was limited to some 100 km, depending on the power of the (mobile) transmitters. In 1946, the first mobile telephone system was installed in the U. S. A. (St. Louis). This system did have an interface to the Public Switched Telephone Network (PSTN), the landline phone system, though this interface was not automated, but rather consisted of human telephone operators. However, with a total of six speech channels for the whole city, the system soon met its limits. This motivated investigations of how the number of users could be increased, even though the allocated spectrum would remain limited. Researchers at AT&T's Bell Labs found the answer: the cellular principle, where the geographical area is divided into cells; different cells might use the same frequencies. To this day, this principle forms the basis for the majority of wireless communications.

Despite the theoretical breakthrough, cellular telephony did not experience significant growth during the 1960s. However, there were exciting developments on a different front: in 1957, the Soviet Union launched the first satellite (Sputnik) and the U. S. A. soon followed. This development fostered research in the new area of satellite communications.

Many basic questions had to be solved, including the effects of propagation through the atmosphere, the impact of solar storms[1], the design of solar panels and other long-lasting energy sources for the satellites, and so on.

To this day, satellite communications is an important area of wireless communications. The most widespread application lies in satellite TV transmission.

3. Analog Cellular Systems

The 1970s saw a revived interest in cellular communications. In scientific research, these years saw the formulation of models for path loss, Doppler spectra, fading statistics, and other quantities that determine the performance of analog telephone systems. A highlight of that work was Jakes' book Microwave Mobile Radio that summed up the state of the art in this area. The 1960s and 1970s also saw a lot of basic research that was originally intended for landline communications, but later also proved to be instrumental for wireless communications. For example, the basics of adaptive equalizers, as well as multicarrier communications, were developed during that time.

For the practical use of wireless telephony, the progress in device miniaturization made the vision of "portable" devices more realistic. Companies like Motorola and AT&T vied for leadership in this area and made vital contributions. Nippon Telephone and Telegraph (NTT) established a commercial cellphone system in Tokyo in 1979. However, it was a Swedish company that built up the first system with large coverage and automated switching: up to that point, Ericsson AB had been mostly known for telephone switches while radio communications was of limited interest to them. However, it was just that expertise in switching technology and the (for that time, daring) decision to use digital switching technology that allowed them to combine different cells in a large area into a single network, and establish the Nordic Mobile Telephone (NMT) system. Note that while the switching technology was digital, the radio transmission technology was still analog, and the systems became therefore known as analog systems. Subsequently, other countries developed their own analog phone standards. The system in the U. S. A., e. g., was called Advanced Mobile Phone System (AMPS).

An investigation of NMT also established an interesting method for estimating market size: business consultants equated the possible number of mobile phone users with the number of Mercedes 600 (the top-of-the-line luxury car at that time) in Sweden. Obviously, mobile telephony could never become a mass market, could it? Similar thoughts must have occurred to the management of the inventor of cellular telephony, AT&T. Upon advice from a consulting company, they decided that mobile telephony could never attract a



significant number of participants and stopped business activities in cellular communications.

The analog systems paved the way for the wireless revolution. During the 1980s, they grew at a frenetic pace and reached market penetrations of up to 10% in Europe, though their impact was somewhat less in the U. S. A.. In the beginning of the 1980s, the phones were "portable", but definitely not handheld. In most languages, they were just called "carphones", because the battery and transmitter were stored in the trunk of the car and were too heavy to be carried around. But at the end of the 1980s, handheld phones with good speech quality and quite acceptable battery lifetime abounded. The quality had become so good that in some markets digital phones had difficulty establishing themselves—there just did not seem to be a need for further improvements.

4. GSM and the Worldwide Cellular Revolution

Even though the public did not see a need for changing from analog to digital, the network operators knew better. Analog phones have a bad spectral efficiency, and due to the rapid growth of the cellular market, operators had a high interest in making room for more customers. Also, research in communications had started its inexorable turn to digital communications, and that included digital wireless communications as well. In the late 1970s and the 1980s, research into spectrally efficient modulation formats, the impact of channel distortions, and temporal variations on digital signals, as well as multiple access schemes and much more, were explored in research labs throughout the world. It thus became clear to the cognoscenti that the real-world systems would soon follow the research.

Again, it was Europe that led the way. The European Telecommunications Standards Institute (ETSI) group started the development of a digital cellular standard that would become mandatory throughout Europe and was later adopted in most parts of the world: Global System for Mobile communications (GSM). The system was developed throughout the 1980s; deployment started in the early 1990s and user acceptance was swift. Due to additional features, better speech quality, and the possibility for secure communications, GSM-based services overtook analog services typically within 2 years of their introduction. In the U. S. A., the change to digital systems was somewhat slower, but by the end of the 1990s, this country also was overwhelmingly digital.

Digital phones turned cellular communications, which was already on the road to success, into a blockbuster. By the year 2000, market penetration in Western Europe and Japan had exceeded 50%, and though the U. S. A.. showed a somewhat delayed development, growth rates were spectacular as well.

The development of wireless systems also made clear the necessity of standards. De-

vices can only communicate if they are compatible, and each receiver can "understand" each transmitter—i. e. , if they follow the same standard. But how should these standards be set? Different countries developed different approaches. The approach in the U. S. A. is "hands-off": allow a wide variety of standards and let the market establish the winner (or several winners). When frequencies for digital cellular communications were auctioned off in the 1990s, the buyers of the spectrum licences could choose the system standard they would use. For this reason, three different standards are now being used in the U. S. A. . A similar approach was used by Japan, where two different systems fought for the market of Second Generation (2G) cellular systems. In both Japan and the U. S. A. , the networks based on different standards work in the same geographical regions, allowing consumers to choose between different technical standards.

The situation was different in Europe. When digital communications were introduced, usually only one operator per country (typically, the incumbent public telephone operators) existed. If each of these operators would adopt a different standard, the result would be high market fragmentation (i. e. , a small market for each standard), without the benefit of competition between operators.

Furthermore, roaming from country to country, which for obvious geographical regions is much more frequent in Europe than in the U. S. A. or Japan, would be impossible. It was thus logical to establish a single common standard for all of Europe.

New Words and Phrases

occurrence *n.* 发生，出现；遭遇，事件
publicize *vt.* 宣布；发表；为……做广告
jungle drum 丛林鼓
cable *n.* 缆绳，电缆
vt. 发电报至；电传
vi. 拍发电报
electromagnetic *adj.* 电磁的
propagation *n.* 传播，传输，蔓延，扩展，波及深度
groundbreaking *adj.* 开创性的，突破性的
noteworthy *adj.* 值得注意的，显著的，重要的
endeavor *vt.* & *vi.* 尝试，试图；尽力，竭力
n. 努力，尽力
bidirectional *adj.* 双向的
optimum *adj.* 最适宜的
n. 最佳效果；最适宜条件
panel *n.* 镶板；面；嵌板；控制板
vt. 把……分格
doppler spectra 多普勒谱
instrumental *adj.* 乐器的；仪器的；有帮助的；起作用的
equalizer *n.* 均衡器，平衡装置；补偿器
miniaturization *n.* 小型化
coverage *n.* 范围，规模
abound *vi.* 丰富，盛产
inexorable *adj.* 不能变更的；不可阻挡的
spectrally *adv.* 幽灵似地，可怕地



mandatory *adj*. 强制的；命令的；受委托的
n. 受托者

blockbuster *n*. 重磅炸弹，了不起的人或事；大片；风靡一时的事物

penetration *n*. 渗透；穿透

spectacular *adj*. 惊人的
n. 壮观的场面，精彩的表演

auction *n*. 拍卖；竞卖；标售
vt. 拍卖；竞卖

fragmentation *n*. 分裂，破碎

Notes

1. solar storms：太阳风暴，指太阳在黑子活动高峰阶段产生的剧烈爆发活动。太阳风暴随太阳黑子活动周期每 11 年发生一次，是一种太阳自身的周期性变化，每个周期内都会有峰年。太阳风暴爆发时释放大量带电粒子所形成的高速粒子流，严重影响地球的空间环境，破坏臭氧层，干扰无线通信，对人体健康也有一定的危害。

Exercises

Ⅰ. Answer the following questions.

1. Why has wireless communications been associated with cellular telephony?
2. What are the two paths to developing new technical solutions?
3. What are the requirements of wireless communications?

Ⅱ. Translate the following sentences into Chinese.

1. For a long time, wireless communications has been associated with cellular telephony, as this is the biggest market segment, and has had the highest impact on everyday lives. In recent times, wireless computer networks have also led to a significant change in working habits and mobility of workers—answering emails in a coffee shop has become an everyday occurrence.

____

____

____

____

____

2. Quite generally, there are two paths to developing new technical solutions: engineering driven and market driven. In the first case, the engineers come up with a brilliant scientific idea—without having an immediate application in mind. As time progresses, the market finds applica-

tions enabled by this idea. In the other approach, the market demands a specific product and the engineers try to develop a technical solution that fulfills this demand.

3. Even though the public did not see a need for changing from analog to digital, the network operators knew better. Analog phones have a bad spectral efficiency, and due to the rapid growth of the cellular market, operators had a high interest in making room for more customers. Also, research in communications had started its inexorable turn to digital communications, and that included digital wireless communications as well.

4. Digital phones turned cellular communications, which was already on the road to success, into a blockbuster. By the year 2000, market penetration in Western Europe and Japan had exceeded 50%, and though the U. S. A. showed a somewhat delayed development, growth rates were spectacular as well.

Abstract Reading

Device-to-device Communications with Wi-Fi Direct: Overview and Experimentation

Wi-Fi Direct is a new technology defined by the Wi-Fi Alliance aimed at enhancing direct device to device communications in Wi-Fi. Thus, given the wide base of devices with Wi-Fi capabilities, and the fact that it can be entirely implemented in software over tradi-



tional Wi-Fi radios, this technology is expected to have a significant impact. In this article we provide a thorough overview of the novel functionalities defined in Wi-Fi Direct, and present an experimental evaluation that portrays the performance to be expected in real scenarios. In particular, our results quantify the delays to be expected in practice when Wi-Fi Direct devices discover each other and establish a connection, and the performance of its novel power saving protocols. To the best of the authors' knowledge this is the first article in the state of the art that provides a wide overview and experimental evaluation of Wi-Fi Direct.

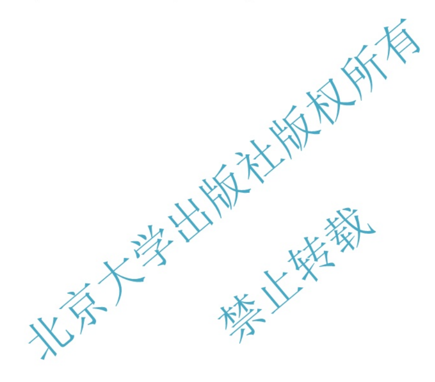
Fractional Frequency Reuse for Interference Management in LTE-advanced Hetnets

Improvement of cell coverage and network capacity are two major challenges for the evolving 4G cellular wireless communication networks such as LTE-Advanced networks. In this context, hierarchical layering of cells with macro base stations coexisting with low-power and shortrange base stations (corresponding to picocells or femtocells) in a service area is considered to be an efficient solution to enhance the spectral efficiency of the network per unit area. Also, such a hierarchical cell deployment, which is referred to as a heterogeneous network, or Het-Net, provides significant improvement in the coverage of indoor and cell edge users and ensures better QoS to the users. Interference mitigation between different layers is one of the key issues that needs to be resolved for successful deployment of HetNets. To this end, fast frequency response, FFR, is considered to be an efficient intercell interference coordination technique for OFDMA-based HetNets. This article focuses on evaluating three state-of-the-art FFR deployment schemes: strict FFR, soft FFR, and FFR-3 schemes for OFDMA-based two-tier HetNets comprising macrocells overlaid with femtocells. Also, a variation of the FFR-3 scheme, which is referred to as the optimal static FFR (OSFFR) scheme, is proposed. A broad comparison among all these FFR schemes is performed by using Monte Carlo simulations considering performance metrics such as outage probability, average network sum rate, and spectral efficiency. Simulation results show that, the average gains in spectral efficiency (bps/Hz) of the network are significantly higher for the proposed scheme when compared to the strict FFR, soft FFR, and FFR-3 schemes.

Integration of Wearable Devices in a Wireless Sensor Network for an E-health Application

Applications based on Wireless Sensor Networks for Internet of Things scenarios are

on the rise. The multiple possibilities they offer have spread towards previously hard to imagine fields, like e-health or human physiological monitoring. An application has been developed for its usage in scenarios where data collection is applied to smart spaces, aiming at its usage in fire fighting and sports. This application has been tested in a gymnasium with real, non-simulated nodes and devices. A Graphic User Interface has been implemented to suggest a series of exercises to improve a sportsman/woman's condition, depending on the context and their profile. This system can be adapted to a wide variety of e-health applications with minimum changes, and the user will interact using different devices, like smart phones, smart watches and/or tablets.



UNIT 13

OPTICAL FIBER COMMUNICATION

Text: Integrated Optics and Photonics

Introduction

This paper contains the discussion of both active and passive optical components and devices which are utilized within optical fiber communications to enable both signal manipulation and processing which facilitates the implementation of high-performance optical fiber communication systems and networks. The integration of such components and devices is also an important and integral aspect to give multiple optical processing functions which are carried out without recourse to conversion back to the electrical domain. Hence the development of integrated optics (IO)[1] for the realization of optical and electro-optical elements integrated together and, more recently, integrated photonics (IP)[2] in which device integration in large numbers can be provided on a single substrate are crucial to the further development of the optical fiber telecommunications network.

In particular, IP refers to the fabrication and integration of several or many components onto a single planar substrate. Such components include beam splitters, couplers, gratings, polarization controllers, interferometers, sources, detectors and optical amplifiers. These components when integrated with planar waveguides constitute the basic building blocks to fabricate more complex planar devices that perform not only optical signal guiding and coupling but also controlling functions such as switching, splitting, multiplexing and demultiplexing of optical signals.

Integrated Optics and Photonics Technologies

Integrated technology for optical devices has developed within optical fiber communi-

cations so that it is now possible to fabricate a complete system onto a single chip. Integration for such devices has become a confluence of several optical or photonic disciplines. Both IO and IP technologies are referred to in the above where the control of the optical devices distinguishes one technology from the other. Electronic control of the optical devices determines the terminology of IO whereas photons control the operation of IP devices. In addition, IP does not involve any optoelectronic conversion of optical signals and hence this technology is also termed as "all-optical". Both IO and IP use planar waveguide technology to provide interconnections between optical components including the basic components for guiding and control of optical signals. IP technology, however, enables the fabrication of subsystems and systems to be realized onto a single substrate. Hence when both active and passive devices are monolithically integrated onto a single substrate in a multilayered integration then these devices are normally referred to as IP devices, while when both active and passive elements fabricated as individual devices are interconnected together they form larger IO devices or circuits. Thus IP can also be seen as a process for the miniaturization and integration of optical systems on a single substrate, and therefore it may be considered as a further enhancement of IO technology, not necessarily as an alternative technology. Both IO and IP seek to provide an alternative to the conversion of an optical signal back into the electrical regime prior to signal processing by allowing such processing to be performed on the optical signal. Hence thin transparent dielectric layers on planar substrates which act as optical waveguides are used in IO and IP to produce small-scale and miniature optical components and circuits. The birth of IO may be traced back to basic ideas outlined by Anderson in 1966. He suggested that a microfabrication technology could be developed for single-mode optical devices with semiconductor and dielectric materials in a similar manner to that which had taken place with electronic circuits. It was in 1969, however, after Miller had introduced the term integrated optics when discussing the long-term outlook in the area, that research began to gain momentum.

Developments in IO have passed the stage where both signal processing and logic functions can be physically realized. Furthermore, such devices may form the building blocks for future digital optical computers. Nevertheless, a number of these advances combine to be closely linked with developments in lightwave communication employing optical fibers.

A major factor in the development of IO is that it is essentially based on single-mode optical waveguides and therefore tends to be incompatible with multimode fiber systems. Hence IO did not make a significant contribution to early deployed optical fiber systems. The advent, however, of single-mode transmission technology further stimulated work in IO in order to provide devices and circuits for these more advanced third-generation systems. Furthermore, the continued expansion of single-mode optical fiber communications



has created a growing market for such IO components. It is also likely that new generations of optical fiber communication systems employing coherent and possibly soliton transmission will lean heavily on IO and IP techniques for their implementation.

The proposals for IO and IP devices and circuits which in many cases involve reinventions of electronic devices and circuits exhibit major advantages other than solely a compatibility with optical fiber communications. Electronic circuits have a practical limitation on speed of operation at a frequency of around 1 010 Hz resulting from their use of metallic conductors to transport electronic charges and build up signals. The large transmission bandwidths (over 100 GHz) currently under investigation for optical fiber communications are already causing difficulties for electronic signal processing within the terminal equipment.

The use of light with its property as an electromagnetic wave of extremely high frequency (1 014 to 1 015 Hz) offers the possibility of high-speed operation around 104 times faster than that conceivable employing electronic circuits. Interaction of light with materials such as semiconductors or transparent dielectrics occurs at speeds in the range of 10－12 (pico) to approaching 10－15 (femto) seconds, thus providing a basis for subpicosecond[3] optical switching.

The other major attribute provided by optical signals interacting within a responsive medium is the ability to utilize lightwaves of different frequencies (or wavelengths) within the same guided wave channel or device. Such frequency division multiplexing allows an information transfer capacity far superior to anything offered by electronics. Moreover, in signal processing terms it facilitates parallel access to information points within an optical system. This possibility for powerful parallel signal processing coupled with ultrahigh speed operation offers tremendous potential for applications within both communications and computing.

The devices of interest in IO and IP are often the counterparts of microwave or bulk optical devices. These include junctions and directional couplers, switches and modulators, filters and wavelength multiplexers, lasers and amplifiers, detectors and bistable elements. It is envisaged that developments in this technology will provide the basis for the next generation of optical networks where full monolithic integration will be used.

The technology associated with the design and fabrication of IP circuits and devices depends upon different factors that mainly result from the characteristics of the substrate material on which the various devices are to be fabricated. The IP process may require serial, parallel or hybrid integration of independent devices. In serial integration of devices, different elements of the optical chip can be interconnected in a consecutive manner and therefore side-, or edge-emitting, or conducting optical devices can be readily integrated on the

same substrate. In the parallel case the chip is constructed by developing columns of devices in which surface- or bottom-emitting devices can effectively be used whereas in hybrid integration IP technology the devices are fabricated using both serial and parallel integration on the same substrate. To gain control of the optical signals, however, additional elements can be developed separately or be directly attached to the IP circuit, or be interconnected to it by optical fibers. In addition, both active and passive devices may be required to be located on the same substrate and therefore hybrid IP integration demands multilayered IP circuits and components to be produced on a single substrate such that they must be compatible with three-dimensional structures of other IO or IP devices.

The enabling technologies for IP mainly rely on silica-on-silicon (SOS) where the waveguide structures comprise three layers, namely the buffer, the core and the cladding. Due to its refractive index match to silica-based optical fiber, vertical light confinement in SOS can be achieved by increasing the refractive index of the core layer relative to the surrounding glass, whereas lateral confinement is obtained through structuring of the core layer. The real benefit of SOS, however, is the ability to apply wafer-scale, planar lithography and processing techniques to integrate substantial numbers of functions either as arrays of identical devices or in the form of customized circuit configurations on single or multiple chips. This integration capability offers an efficient platform for the implementation of typical fiber-based functionalities such as optical power splitters or combiners, couplers, wavelength-selective couplers, multiplexers/demultiplexers and optical gain elements. Furthermore, optical switches and controllable attenuators based on the thermo-optic effect can also be fabricated. In addition, some devices can be fabricated using a silicon-on-liquid (SOL) gel process which cannot be implemented using SOS techniques. The SOL gel process is a versatile solution-based technique for making ceramic and glass materials which involves the transition of a system from a liquid (or solution) into a gel. Applying the SOL gel and SOS techniques, it is possible to fabricate thin-film coatings, ceramic fibers and waveguide based optical amplifiers. Further to the above integration technologies used for IP devices, a silicon-on-insulator (SOI) approach is also used to produce microwave guide bends and couplers at a reduced scale while maintaining compatibility with the standard silicon fabrication techniques. SOI based on CMOS technology that has revolutionized both electronic and optoelectronic integrated circuit technologies is now usefully applied as an IP technology. Using this integration technique, active components like optical sources, detectors and amplifiers can be coupled to other IP devices. Furthermore, the more recently developed photonic crystal waveguide technology is also compatible with this integration technology.



New Words and Phrases

amplification *n*. 放大；放大率
facilitate *vt*. 促进，助长；使容易；帮助
beam splitter 分束器
coupler *n*. 耦合器；连接器
grating *n*. 光栅；格栅
polarization controller 偏振控制器
interferometer *n*. 干涉仪
amplifier *n*. 放大器
fabricate *vt*. 制造；装配；捏造
demultiplexing *n*. 多路解编
photonics *n*. 光子学
confluence *n*. 合流；融合
terminology *n*. 术语，术语学
optoelectronic *adj*. 光电子的
subsystem *n*. 子系统，分系统
monolithically *adv*. 单片
multilayered *adj*. 多层的
microfabrication *n*. 微细加工
semiconductor *n*. 半导体
advent *n*. 出现，到来；将临期
soliton *n*. 孤子；光孤子
reinvention *n*. 再创造、重塑
solely *adv*. 唯一地；仅仅
compatibility *n*. 适合；互换性；通用性
responsive *adj*. 响应的；有求必应；应叫性
ultrahigh *adj*. 超高的，特高的
bulk *n*. (大)体积
  adj. 大批的，大量的
filter *n*. 滤波器
multiplexer *n*. 多路调制器
bistable *adj*. 双稳态的
envisage *vt*. 想象，设想；正视，面对
monolithic *n*. 单片；整体式
serial *n*. 串口；序列号；串行
  adj. 连续的
consecutive *adj*. 连续的，连贯的
emit *vt*. 发出；发射；颁布
fabricated *adj*. 编造的
  vt. 编造(fabricate 的过去分词)
buffer *n*. 缓冲器；缓冲区
  vt. 缓冲
cladding *n*. 包层；电镀；包壳；覆盖层
refractive *adj*. 折射的
silica *n*. 硅石，二氧化硅
vertical *adj*. 垂直的，竖立的
confinement *n*. 监禁，关押；分娩
lithography *n*. 光刻；平版印刷
customize *vt*. 定制
attenuator *n*. 衰减器
ceramic *adj*. 陶器的
  n. 陶瓷制品；陶瓷器
bend *vt*. & *vi*. 使弯曲；弯管；弯曲变形；弯
compatibility *n*. 适合；互换性；通用性

Notes

1. integrated optics (IO)：集成光学。集成光学是研究媒质薄膜中的光学现象以及光学元件集成化的一门学科。它是在激光技术发展过程中，由于光通信、光学信息处理等的

需要，而逐步形成和发展起来的。

2. integrated photonics (IP)：集成光子学，主要研究光子集成技术，即将光电子器件和电子器件集成在芯片上。

3. subpicosecond：亚皮秒。1 秒＝1 000 毫秒＝1 000 000 微秒＝1 000 000 000 纳秒＝1 000 000 000 000 皮秒；1 皮秒＝一万亿分之一秒。亚皮秒就是比皮秒还要小的时间尺度。

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. optical fiber communication
2. integrated optics
3. electro-optical element
4. wafer-scale
5. thermo-optic effect
6. waveguide based optical amplifier
7. integrated photonics
8. soliton transmission

Ⅱ. Fill in the blanks with missing word(s) from the table below.

optical	confining	transmission	microelectronics
circuits	congestion	detected	hazardous
diameters	expansion	fabricated	visible
deposit	transfer	applications	extension
copper	interface	passive	noninvasive

1. The use of ________ optical carrier waves or light for communication has been common for many years. Simple systems such as signal fires, reflecting mirrors and, more recently, signaling lamps have provided successful, if limited, information ________.

2. The use of circular dielectric waveguide structures for ________ light is universally utilized within optical fiber communications. Both IO and IP involve an ________ of this guided wave optical technology through the use of planar ________ waveguides to confine and guide the light in guided wave devices and circuits.

3. Planar waveguide structures are produced using several different techniques which have in large part been derived from the ________ industry. For example, ________ devices may be fab-



ricated by radio-frequency sputtering to ________ thin films of glass onto glass substrates.

4. Optical fibers have very small ________ which are often no greater than the diameter of a human hair. Hence, even when such fibers are covered with protective coatings they are far smaller and much lighter than corresponding ________ cables. This is a tremendous boon towards the alleviation of duct ________ in cities, as well as allowing for an ________ of signal ________ within mobiles such as aircraft, satellites and even ships.

5. Optical fibers which are ________ from glass, or sometimes a plastic polymer, are electrical insulators and therefore, unlike their metallic counterparts, they do not exhibit earth loop and ________ problems. Furthermore, this property makes optical fiber transmission ideally suited for communication in electrically ________ environments as the fibers create no arcing or spark hazard at abrasions or short ________.

6. The light from optical fibers does not radiate significantly and therefore they provide a high degree of signal security. Unlike the situation with copper cables, a transmitted optical signal cannot be obtained from a fiber in a ________ manner (i. e. , without drawing optical power from the fiber). Therefore, in theory, any attempt to acquire a message signal transmitted optically may be ________. This feature is obviously attractive for military, banking and general data transmission (i. e. , computer network) ________.

Ⅲ. Translate the following sentences into Chinese.

1. Communication may be broadly defined as the transfer of information from one point to another. When the information is to be conveyed over any distance a communication system is usually required. Within a communication system the information transfer is frequently achieved by superimposing or modulating the information onto an electromagnetic wave which acts as a carrier for the information signal. This modulated carrier is then transmitted to the required destination where it is received and the original information signal is obtained by demodulation. Sophisticated techniques have been developed for this process using electromagnetic carrier waves operating at radio frequencies as well as microwave and millimeter wave frequencies. However, "communication" may also be achieved using an electromagnetic carrier which is selected from the optical range of frequencies.

2. Communication using an optical carrier wave guided along a glass fiber has a number of extremely attractive features, several of which were apparent when the technique was originally conceived. Furthermore, the advances in the technology to date have surpassed even the most optimistic predictions, creating additional advantages. Hence it is useful to consider the merits and special features offered by optical fiber communications over more conventional electrical communications.

Reading: Coherent and Phase-modulated

The direct detection of an intensity-modulated optical carrier is basically a photon counting process where each detected photon is converted into an electron-hole pair (or, in the case of the APD, a number of pairs due to avalanche gain).

Conventional direct detection receivers, however, are generally limited by noise generated in the detector and preamplifier except at very high signal-to-noise ratios (SNRs). The sensitivity of such square-law detection systems is therefore reduced below the fundamental quantum noise limit by at least 10 to 20 dB. This is particularly the case at longer wavelengths (i. e., 1.3 to 1.6μm) and at higher transmission rates since the electronic preamplifier usually has a rising input optical power with frequency requirement. For a good APD receiver operating in the wavelength range 1.3 to 1.6μm this corresponds to between 700 and 1 000 photons per bit required to maintain a bit-error-rate (BER) of 10^{-9}.

Improvements in receiver sensitivity, together with wavelength selectivity, may be obtained using the well-known coherent detection techniques (i. e., heterodyne and homodyne detection) for the optical signal. Unlike direct detection in which the optical signal is converted directly into a demodulated electrical output, such coherent optical receivers first add to the incoming optical signal from a locally generated optical wave prior to detecting the sum. The resulting photocurrent is a replica of the original signal which is translated down in frequency from the optical domain (around 105 GHz) to the radio domain (up to several GHz) and where conventional electronic techniques can be used for further signal



processing and demodulation. Hence an ideal coherent receiver operating in the 1.3 to 1.6μm wavelength region requires a signal energy of only 10 to 20 photons per bit to achieve a BER of 10^{-9}. Hence coherent detection potentially provides a substantial benefit for high-speed systems operating at longer wavelengths. A possible improvement in receiver sensitivity using conventional coherent detection of up to 20 dB can be obtained over direct detection. Furthermore, such enhanced receiver sensitivity could translate into increases in repeater spacings of the order of 100 km or more when using low-loss fiber at a wavelength of 1.55μm. Hence, the improved sensitivity of 5 to 20 dB which results from the photomixing gain in the coherent receiver could provide:

(1) Increased repeater spacings for both inland and undersea transmission systems.

(2) Higher transmission rates over existing routes without reducing repeater spacings.

(3) Increased power budgets to compensate for losses associated with couplers and optical multiplexer/demultiplexer devices in distribution networks.

(4) Improved sensitivity to optical test equipment such as optical time domain reflectometers.

Although possible increases in the transmission distance between repeaters created the initial impetus for the pursuit of coherent transmission within optical fiber communications, by 1990, before the widespread deployment of optical amplifiers, it was also perceived that coherent techniques would enable a further massive step to be taken in the exploitation of the transmission capacity of optical fiber systems. This improvement would be facilitated by the inherent wavelength selectivity afforded by the coherent receiver allowing efficient access to the vast optical bandwidth available in single-mode fibers. For example, it was suggested for the low-loss window between 1.3 and 1.6μm being over 50 000 GHz that coherent transmission would permit wavelength division multiplexing of huge channel numbers with channel spacings of only a few hundred megahertz.

The successful introduction into the optical telecommunication network of erbium-doped fiber amplifiers in the early 1990s, however, caused the pursuit of coherent transmission to be virtually discontinued as both improved direct detection receiver sensitivity through amplification and hence increased transmission distance together with dense wavelength division multiplexing with channel spacings of 50 GHz (0.4 nm) or even 12.5 GHz (0.1 nm) could be facilitated.

The modulation formats that may be employed within coherent optical fiber communications are essentially the same as those used in coherent electrical line and radio communications. Modulation formats of this type were discussed in later parts in a slightly different context, namely the generation of subcarriers for electrical frequency division multiplexing prior to intensity modulation of the optical source. In these cases direct detection of the op-

tical signal is carried out at the receiver with subsequent electrical demodulation for the subcarriers. Such systems only provide improvements in the SNR over baseband IM/DD systems at the expense of a substantial bandwidth penalty. When a narrow-linewidth injection laser (less than 1 MHz) is used in an optical fiber communication system, however, it is possible to directly modulate the coherent optical carrier in amplitude (direct AM), frequency (direct FM) and phase (direct PM) prior to demodulation using a coherent optical receiver. In the case of digital transmission this implies amplitude, frequency or phase shift keying (i. e. , ASK, FSK or PSK) modulation techniques.

Direct modulation for coherent optical fiber transmission is provided using an external modulator which is fed by a semiconductor laser source operating in continuous wave mode. In electrical digital communications, coherent demodulation requires recovery of the carrier frequency signal through either heterodyne or homodyne detection. Coherent optical communications, however, has had a wider usage of the terminology. An optical fiber communication system was referred to as coherent when there was optical signal mixing even though carrier recovery may not occur. Hence an optical coherent system could employ a demodulator that did not perform carrier recovery but used noncoherent or envelope detection. In this context differential phase shift keying (DPSK) is considered a noncoherent electrical modulation technique but it has generally been referred to as a coherent optical communication modulation format. Moreover, a coherent optical receiver is referred to as synchronous or asynchronous (i. e. , nonsynchronous) depending upon whether it operates with or without phase tracking, respectively. The latter asynchronous receiver normally employs power or envelope detection.

Although, as indicated above, research in coherent optical fiber communications effectively ceased in the early 1990s as a result of the successful introduction of fiberamplifiers, DPSK using, in particular, asynchronous detection started to receive renewed interest following specific experimental demonstrations in 2002. In the context of the renewed focus these high-performance, long-haul systems with return-to-zero DPSK transmission, where a phase-to-intensity conversion takes place at the receiver prior to a direct detection process, have subsequently been widely demonstrated. As optical mixing with an independent optical signal does not necessarily take place at the receiver to achieve carrier recovery, such systems tend to be referred to as being optically phase-modulated or phase shift keyed, or when appropriate, as self-coherent.

The term coherent is then often utilized for the case of coherent detection in which the phase reference at the receiver is typically provided by a local oscillator laser that beats with the received optical signal to produce constructive and destructive interference. There is an increasing adoption of this stricter approach to the use of the terminology and there-



fore the paper title reflects this trend by incorporating both the "coherent" and "phase-modulated" system terms in order to encompass systems using advanced modulation formats with direct detection, and also to emphasize the growth of phase modulation as the currently preferred advanced optical modulation format in competition with intensity modulation.

While prior to the 1990s a major factor in the pursuit of coherent and phase-modulated optical fiber systems had been the perceived improvement in receiver sensitivities to enable transmission over longer distances, a main focus of the renewed interest since 2002 has been to increase spectral efficiency. Moreover, direct detection phase-modulated optical fiber systems do also provide for a sensitivity gain over IM/DD systems, albeit at more modest levels than when using synchronous detection. In particular, it should be noted that the theoretical sensitivity gains which can be determined for both PSK and DPSK over IM/DD are substantially reduced when optical amplifiers are deployed within the latter systems.

New Words and Phrases

coherent *adj*. 一致的；连贯的
photon *n*. 光子，光量子
preamplifier *n*. （无线电）前置放大器
quantum *n*. 量子；定量
heterodyne *n*. 外差；外差法；外差振荡器
homodyne *n*. 同色异构；同步检波
demodulated *adj*. 已解调的
photocurrent *n*. 光电流
replica *n*. 复制品
reflectometer *n*. 反射计
impetus *n*. 动力；促进；势头；声势
exploitation *n*. 开发；利用；剥削；广告推销
erbium-doped *n*. 铒掺杂（半导体工艺中的一种掺杂技术）
subcarriers *n*. 子载波
penalty *n*. 惩罚；刑罚；害处；足球点球
linewidth *n*. 行距；行宽；线幅；线宽
injection *n*. 注入；注射剂
asynchronous *adj*. 异步的
fiberamplifier *n*. 光纤放大器；光放大器
renewed *adj*. 更新的；重建的
vt. & *vi*. （使）复原，（使）更新（venew 的过去分词）
long-haul *adj*. 长途的
encompass *vt*. 围绕，包围；包含或包括某事物；完成
deploy *vt*. & *vi*. 使展开；施展；有效地利用

Exercises

Ⅰ. Answer the following questions.

1. What's the direct detection of an intensity-modulated optical carrier?

2. What could the improved sensitivity of 5 to 20 dB provide?

3. What are the common phase shift keying methods?

Ⅱ. Translate the following sentences into Chinese.

1. Improvements in receiver sensitivity, together with wavelength selectivity, may be obtained using the well-known coherent detection techniques (i. e. , heterodyne and homodyne detection) for the optical signal. Unlike direct detection in which the optical signal is converted directly into a demodulated electrical output, such coherent optical receivers first add to the incoming optical signal from a locally generated optical wave prior to detecting the sum.

2. Direct modulation for coherent optical fiber transmission is provided using an external modulator which is fed by a semiconductor laser source operating in continuous wave mode. In electrical digital communications, coherent demodulation requires recovery of the carrier frequency signal through either heterodyne or homodyne detection.

3. The direct detection of an intensity-modulated optical carrier is basically a photon counting process where each detected photon is converted into an electron-hole pair (or, in the case of the APD, a number of pairs due to avalanche gain).



Abstract Reading

Optical Memory Made of Photonic Crystal Working over the C-band of ITU

After several decades pushing the technology and the development of the world the electronics is giving space for technologies that use light. We propose and analyze an optical memory embedded in a nonlinear Photonic Crystal (PhC), whose system of writing and reading of data is controlled by external optical command signals. This optical memory is based on two optical Nonlinear Directional Couplers (NDCs), which are connected to a shared ring. These NDCs have a small coupling length and the optical memory can works in an all-optical broadband telecommunications system (C-Band of ITU (International Telecommunication Union)—wavelength from 1 530 nm to 1 565 nm).

Characterization and Training of a 19-element Piezoelectric Deformable Mirror for Lensing

Characterizing the fundamental response and operational parameters of a deformable mirror is a critical first step in the design of an adaptive optics system. This paper describes the characterization and training of a piezoelectric deformable mirror (PDM) to be implemented on a low order Adaptive Optics (AO) system at Short Wave Infrared (SWIR) wavelengths for free-space optical communications systems. The data were analyzed using commercial and customized software.

Algorithmical Analysis of Information-theoretic Aspects of Secure Communication over Optical-fiber Quantum Channels

The information-theoretic security of optical-fiber based quantum communication is the fundamental question of quantum cryptography. Quantum cryptographic schemes use photons as information carriers. The physical properties of photons make it possible to use quantum bits to realize unconditionally secure quantum communication over the current standard optical fiber network. Quantum cryptography is one of the most important and advanced fields in the area of quantum information processing. This paper analyzes the information-theoretic security of the most important and prevalent optical-fiber based QKD

schemes, such as BB84, Six-state and DPS QKD schemes, using efficient information geometric approaches. We study the information-theoretic impacts of the most general eavesdropping attacks against these protocols using efficient algorithms. Currently, the ability to perform these attacks is well beyond today's technological capabilities; however, in the future, these types of attacks can be used to eavesdrop on quantum communications over optical fibers. The information-theoretic security of these protocols is analyzed by information geometric algorithms and abstract geometrical objects. To describe the security of the protocols, we introduce the quantum informational ball representation, and we discover the connection between the length of the optical fiber and the radius of the quantum informational ball. For practical reasons, we will also demonstrate our algorithm for the DPS QKD protocol.

UNIT 14

COMPUTER NETWORKS

Text: Business Applications of Computer Networks

Most companies have a substantial number of computers. For example, a company may have a computer for each worker and use them to design products, write brochures, and do the payroll. Initially, some of these computers may have worked in isolation from the others, but at some point, management may have decided to connect them to be able to distribute information throughout the company.

Put in slightly more general form, the issue here is resource sharing. The goal is to make all programs, equipment, and especially data available to anyone on the network without regard to the physical location of the resource or the user. An obvious and widespread example is having a group of office workers share a common printer. None of the individuals really needs a private printer, and a high-volume networked printer is often cheaper, faster, and easier to maintain than a large collection of individual printers.

However, probably even more important than sharing physical resources such as printers, and tape backup systems, is sharing information. Companies small and large are vitally dependent on computerized information. Most companies have customer records, product information, inventories, financial statements, tax information, and much more online. If all of its computers suddenly went down, a bank could not last more than five minutes. A modern manufacturing plant, with a computer-controlled assembly line, would not last even 5 seconds. Even a small travel agency or three-person law firm is now highly dependent on computer networks for allowing employees to access relevant information and documents instantly.

For smaller companies, all the computers are likely to be in a single office or perhaps a single building, but for larger ones, the computers and employees may be scattered over

dozens of offices and plants in many countries. Nevertheless, a sales person in New York might sometimes need access to a product inventory database in Singapore. Networks called VPNs (Virtual Private Networks) may be used to join the individual networks at different sites into one extended network. In other words, the mere fact that a user happens to be 15 000 km away from his data should not prevent him from using the data as though they were local. This goal may be summarized by saying that it is an attempt to end the "tyranny of geography".

In the simplest of terms, one can imagine a company's information system as consisting of one or more databases with company information and some number of employees who need to access them remotely. In this model, the data are stored on powerful computers called servers. Often these are centrally housed and maintained by a system administrator. In contrast, the employees have simpler machines, called clients, on their desks, with which they access remote data, for example, to include in spreadsheets they are constructing. (Sometimes we will refer to the human user of the client machine as the "client", but it should be clear from the context whether we mean the computer or its user.) The client and server machines are connected by a network. Note that we have shown the network as a simple oval, without any detail. We will use this form when we mean a network in the most abstract sense. When more detail is required, it will be provided. This whole arrangement is called the client-server model. It is widely used and forms the basis of much network usage. The most popular realization is that of a Web application, in which the server generates Web pages based on its database in response to client requests that may update the database. The client-server model is applicable when the client and server are both in the same building (and belong to the same company), but also when they are far apart. For example, when a person at home accesses a page on the World Wide Web, the same model is employed, with the remote Web server being the server and the user's personal computer being the client. Under most conditions, one server can handle a large number (hundreds or thousands) of clients simultaneously.

If we look at the client-server model in detail, we see that two processes (i. e. , running programs) are involved, one on the client machine and one on the server machine. Communication takes the form of the client process sending a message over the network to the server process. The client process then waits for a reply message. When the server process gets the request, it performs the requested work or looks up the requested data and sends back a reply. A second goal of setting up a computer network has to do with people rather than information or even computers. A computer network can provide a powerful communication medium among employees. Virtually every company that has two or more computers now has email (electronic mail), which employees generally use for a great deal



of daily communication. In fact, a common gripe around the water cooler is how much email everyone has to deal with, much of it quite meaningless because bosses have discovered that they can send the same (often content-free) message to all their subordinates at the push of a button. Telephone calls between employees may be carried by the computer network instead of by the phone company. This technology is called IP telephony or Voice over IP (VoIP) [1] when Internet technology is used. The microphone and speaker at each end may belong to a VoIP-enabled phone or the employee's computer.

Companies find this a wonderful way to save on their telephone bills. Other, richer forms of communication are made possible by computer networks. Video can be added to audio so that employees at distant locations can see and hear each other as they hold a meeting. This technique is a powerful tool for eliminating the cost and time previously devoted to travel. Desktop sharing lets remote workers see and interact with a graphical computer screen. This makes it easy for two or more people who work far apart to read and write a shared blackboard or write a report together. When one worker makes a change to an online document, the others can see the change immediately, instead of waiting several days for a letter. Such a speedup makes cooperation among far-flung groups of people easy where it previously had been impossible. More ambitious forms of remote coordination such as telemedicine are only now starting to be used (e.g., remote patient monitoring) but may become much more important. It is sometimes said that communication and transportation are having a race, and whichever wins will make the other obsolete.

A third goal for many companies is doing business electronically, especially with customers and suppliers. This new model is called e-commerce (electronic commerce) and it has grown rapidly in recent years. Airlines, bookstores, and other retailers have discovered that many customers like the convenience of shopping from home. Consequently, many companies provide catalogs of their goods and services online and take orders online. Manufacturers of automobiles, aircraft, and computers, among others, buy subsystems from a variety of suppliers and then assemble the parts. Using computer networks, manufacturers can place orders electronically as needed. This reduces the need for large inventories and enhances efficiency.

New Words and Phrases

brochure *n.* 小册子，手册

payroll *n.* 工资单

tyranny *n.* 暴虐；专横

spreadsheet *n.* 电子表格

oval *adj.* 椭圆形的
n. 椭圆形

subordinate *adj.* 级别或职位较低的；下级的；次要的；附属的

n. 部属；部下，下级
vt. 使……居下位，使在次级；使服从；使从属

far-flung *adj.* 遥远的，广泛的

telemedicine *n.* （通过遥测、电话、电视等手段求诊的）远距离医学

obsolete *adj.* 废弃的；老式的
n. 废词；陈腐的人
vt. 淘汰；废弃

Notes

1. VoIP：Voice over Internet Protocol，简而言之就是将模拟信号(Voice)数字化，以数据封包(Data Packet)的形式在 IP 网络(IP Network)上实时传递。VoIP 最大的优势是能广泛地采用 Internet 和全球 IP 互连的环境，提供比传统业务更多、更好的服务。VoIP 可以在 IP 网络上便宜地传送语音、传真、视频和数据等业务，如统一消息业务、虚拟电话、虚拟语音/传真邮箱、查号业务、Internet 呼叫中心、Internet 呼叫管理、电话视频会议、电子商务、传真存储转发和各种信息的存储转发等。

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. high-volume networked printer
2. tape backup
3. assembly line
4. Virtual Private Networks
5. client-server
6. e-commerce
7. telemedicine

Ⅱ. Fill in the blanks with missing word(s) from the table below.

subnet	behalf	determines	telecommunications
datagrams	optical	representing	components
centralized	destination	emphasize	implemented
independent	layer	allocation	hierarchy
protocols	allocation	transmitted	transmission

1. Dynamic ________ methods for a common channel are either ________ or decentralized. In the centralized channel ________ method, there is a single entity, for example, the base station in cellular networks, which ________ who goes next.

2. In most WANs, the ________ consists of two distinct ________: transmission lines and switching elements. ________ lines move bits between machines. They can be made of copper wire, ________ fiber, or even radio links. Most companies do not have transmission lines lying about, so instead they lease the lines from a ________ company.

3. Services and ________ are distinct concepts. This distinction is so important that we ________ it again here. A service is a set of primitives (operations) that a layer provides to the layer above it. The service defines what operations the ________ is prepared to perform on ________ of its users, but it says nothing at all about how these operations are ________.

4. Information can be ________ on wires by varying some physical property such as voltage or current. By ________ the value of this voltage or current as a single-valued function of time, $f(t)$, we can model the behavior of the signal and analyze it mathematically.

5. Together with the network layer, the transport layer is the heart of the protocol ________. The network layer provides end-to-end packet delivery using ________ or virtual circuits. The transport layer builds on the network layer to provide data transport from a process on a source machine to a process on a ________ machine with a desired level of reliability that is ________ of the physical networks currently in use.

Ⅲ. Translate the following sentences into Chinese.

1. PANs (Personal Area Networks) let devices communicate over the range of a person. A common example is a wireless network that connects a computer with its peripherals. Almost every computer has an attached monitor, keyboard, mouse, and printer. Without using wireless, this connection must be done with cables. So many new users have a hard time finding the right cables and plugging them into the right little holes (even though they are usually color coded) that most computer vendors offer the option of sending a technician to the user's home to do it.

____

____

____

____

____

2. It is worth spending a little more time discussing LANs in the home. In the future, it is likely that every appliance in the home will be capable of communicating with every

other appliance, and all of them will be accessible over the Internet. This development is likely to be one of those visionary concepts that nobody asked for (like TV remote controls or mobile phones), but once they arrived nobody can imagine how they lived without them.

____
____
____
____
____

3. To reduce their design complexity, most networks are organized as a stack of layers or levels, each one built upon the one below it. The number of layers, the name of each layer, the contents of each layer, and the function of each layer differ from network to network. The purpose of each layer is to offer certain services to the higher layers while shielding those layers from the details of how the offered services are actually implemented. In a sense, each layer is a kind of virtual machine, offering certain services to the layer above it.

____
____
____
____
____

Reading: Home Applications

In 1977, Ken Olsen was president of the Digital Equipment Corporation, then the number two computer vendor in the world (after IBM). When asked why Digital was not going after the personal computer market in a big way, he said: "There is no reason for any individual to have a computer in his home." History showed otherwise and Digital no longer exists. People initially bought computers or word processing and games. Recently, the biggest reason to buy a home computer was probably for Internet access. Now, many consumer electronic devices, such as set-top boxes, game consoles, and clock radios, come with embedded computers and computer networks, especially wireless networks, and home networks are broadly used for entertainment, including listening to, looking at, and creating music, photos, and videos.

Internet access provides home users with connectivity to remote computers. As with

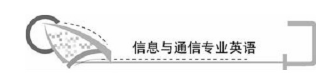


companies, home users can access information, communicate with other people, and buy products and services with e-commerce. The main benefit now comes from connecting outside of the home. Bob Metcalfe, the inventor of Ethernet, hypothesized that the value of a network is proportional to the square of the number of users because this is roughly the number of different connections that may be made (Gilder, 1993). This hypothesis is known as "Metcalfe's law". It helps to explain how the tremendous popularity of the Internet comes from its size.

Access to remote information comes in many forms. It can be surfing the World Wide Web for information or just for fun. Information available includes the arts, business, cooking, government, health, history, hobbies, recreation, science, sports, travel, and many others. Fun comes in too many ways to mention, plus some ways that are better left unmentioned.

Many newspapers have gone online and can be personalized. For example, it is sometimes possible to tell a newspaper that you want everything about corrupt politicians, big fires, scandals involving celebrities, and epidemics, but no football, thank you. Sometimes it is possible to have the selected articles downloaded to your computer while you sleep. As this trend continues, it will cause massive unemployment among 12-year-old paperboys, but newspapers like it because distribution has always been the weakest link in the whole production chain. Of course, to make this model work, they will first have to figure out how to make money in this new world, something not entirely obvious since Internet users expect everything to be free.

The next step beyond newspapers (plus magazines and scientific journals) is the online digital library. Many professional organizations, such as the ACM (www. acm. org) and the IEEE Computer Society (www. computer. org), already have all their journals and conference proceedings online. Electronic book readers and online libraries may make printed books obsolete. Skeptics should take note of the effect the printing press had on the medieval illuminated manuscript. Much of this information is accessed using the client-server model, but there is different, popular model for accessing information that goes by the name of peer-to-peer communication). In this form, individuals who form a loose group can communicate with others in the group. Every person can, in principle, communicate with one or more other people; there is no fixed division into clients and servers. Many peer-to-peer systems, such BitTorrent (Cohen, 2003), do not have any central database of content. Instead, each user maintains his own database locally and provides a list of other nearby people who are members of the system. A new user can then go to any existing member to see what he has and get the names of other members to inspect for more content and more names. This lookup process can be repeated indefinitely to build up a large local



database of what is out there.

It is an activity that would get tedious for people but computers excel at it. Peer-to-peer communication is often used to share music and videos. It really hit the big time around 2000 with a music sharing service called Napster that was shut down after what was probably the biggest copyright infringement case in all of recorded history (Lam and Tan, 2001; and Macedonia, 2000). Legal applications for peer-to-peer communication also exist. These include fans sharing public domain music, families sharing photos and movies, and users downloading public software packages. In fact, one of the most popular Internet applications of all, email, is inherently peer-to-peer. This form of communication is likely to grow considerably in the future.

All of the above applications involve interactions between a person and a remote database full of information. The second broad category of network use is person-to-person communication, basically the 21st century's answer to the 19th century's telephone. E-mail is already used on a daily basis by millions of people all over the world and its use is growing rapidly. It already routinely contains audio and video as well as text and pictures. Smell may take a while. Any teenager worth his or her salt is addicted to instant messaging. This facility, derived from the UNIX talk program in use since around 1970, allows two people to type messages at each other in real time. There are multi-person messaging services too, such as the Twitter service that lets people send short text messages called "tweets" to their circle of friends or other willing audiences.

The Internet can be used by applications to carry audio (e. g. , Internet radio stations) and video (e. g. , YouTube). Besides being a cheap way to call to distant friends, these applications can provide rich experiences such as telelearning, meaning attending 8 A. M. classes without the inconvenience of having to get out of bed first. In the long run, the use of networks to enhance human-to-human communication may prove more important than any of the others. It may become hugely important to people who are geographically challenged, giving them the same access to services as people living in the middle of a big city.

Between person-to-person communications and accessing information are social network applications. Here, the flow of information is driven by the relationships that people declare between each other. One of the most popular social networking sites is Facebook. It lets people update their personal profiles and shares the updates with other people who they have declared to be their friends.

Other social networking applications can make introductions via friends of friends, send news messages to friends such as Twitter above, and much more. Even more loosely, groups of people can work together to create content. A wiki, for example, is a collaborative Web site that the members of a community edit. The most famous wiki is the Wikipe-



dia, an encyclopedia anyone can edit, but there are thousands of other wikis.

Our third category is electronic commerce in the broadest sense of the term. Home shopping is already popular and enables users to inspect the online catalogs of thousands of companies. Some of these catalogs are interactive, showing products from different viewpoints and in configurations that can be personalized. After the customer buys a product electronically but cannot figure out how to use it, online technical support may be consulted.

Another area in which e-commerce is widely used is access to financial institutions. Many people already pay their bills, manage their bank accounts, and handle their investments electronically. This trend will surely continue as networks become more secure.

One area that virtually nobody foresaw is electronic flea markets (e-flea). Online auctions of second-hand goods have become a massive industry. Unlike traditional e-commerce, which follows the client-server model, online auctions are peer-to-peer in the sense that consumers can act as both buyers and sellers. Some of these forms of e-commerce have acquired cute little tags based on the fact that "to" and "2" are pronounced the same.

Our fourth category is entertainment. This has made huge strides in the home in recent years, with the distribution of music, radio and television programs, and movies over the Internet beginning to rival that of traditional mechanisms. Users can find, buy, and download MP3 songs and DVD-quality movies and add them to their personal collection. TV shows now reach many homes via IPTV (IP TeleVision) systems that are based on IP technology instead of cable TV or radio transmissions. Media streaming applications let users tune into Internet radio stations or watch recent episodes of their favorite TV shows. Naturally, all of this content can be moved around your house between different devices, displays and speakers, usually with a wireless network.

Soon, it may be possible to search for any movie or television program ever made, in any country, and have it displayed on your screen instantly. New films may become interactive, where the user is occasionally prompted for the story direction (should Macbeth murder Duncan or just bide his time?) with alternative scenarios provided for all cases. Live television may also become interactive, with the audience participating in quiz shows, choosing among contestants, and so on.

Another form of entertainment is game playing. Already we have multiperson real-time simulation games, like hide-and-seek in a virtual dungeon, and flight simulators with the players on one team trying to shoot down the players on the opposing team. Virtual worlds provide a persistent setting in which thousands of users can experience a shared reality with three-dimensional graphics.

Our last category is ubiquitous computing, in which computing is embedded into eve-

ryday life, as in the vision of Mark Weiser (1991). Many homes are already wired with security systems that include door and window sensors, and there are many more sensors that can be folded in to a smart home monitor, such as energy consumption. Your electricity, gas and water meters could also report usage over the network. This would save money as there would be no need to send out meter readers. And your smoke detectors could call the fire department instead of making a big noise (which has little value if no one is home). As the cost of sensing and communication drops, more and more measurement and reporting will be done with networks.

Increasingly, consumer electronic devices are networked. For example, some high-end cameras already have a wireless network capability and use it to send photos to a nearby display for viewing. Professional sports photographers can also send their photos to their editors in real-time, first wirelessly to an access point then over the Internet. Devices such as televisions that plug into the wall can use power-line networks to send information throughout the house over the wires that carry electricity. It may not be very surprising to have these objects on the network, but objects that we do not think of as computers may sense and communicate information too. For example, your shower may record water usage, give you visual feedback while you lather up, and report to a home environmental monitoring application when you are done to help save on your water bill.

A technology called RFID (Radio Frequency Identification) will push this idea even further in the future. RFID tags are passive (i.e., have no battery) chips the size of stamps and they can already be affixed to books, passports, pets, cred it cards, and other items in the home and out. This lets RFID readers locate and communicate with the items over a distance of up to several meters, depending on the kind of RFID. Originally, RFID was commercialized to replace barcodes. It has not succeeded yet because barcodes are free and RFID tags cost a few cents. Of course, RFID tags offer much more and their price is rapidly declining. They may turn the real world into the Internet of things (ITU, 2005).

New Words and Phrases

set-top *n*. 机顶
console *n*. 控制台；操纵台
hypothesize *vt*. 假设，假定
Metcalfe's law 梅特卡夫定律
scandal *n*. 丑闻；流言蜚语；诽谤；公愤
epidemic *adj*. 流行的；传染性的
obsolete *adj*. 废弃的；老式的
vt. 淘汰；废弃
skeptic *n*. 怀疑论者；怀疑者
medieval *adj*. 中世纪的；原始的
peer-to-peer *n*. 点对点
illuminate *vt*. 阐明，说明
manuscript *n*. 手稿；原稿
inherently *adv*. 内在地；固有地；天性地



derive *vt*. 源于；得自
collaborative *adj*. 合作的，协作的
configuration *n*. 配置；结构；外形
auction *vt*. 拍卖；竞卖
n. 拍卖
stride *n*. 大步；步幅；进展
vt. 跨过；大踏步走过
contestant *n*. 竞争者；争辩者
simulation *n*. 仿真；模拟
ubiquitous *adj*. 普遍存在的；无所不在的
affix *vt*. 粘上；署名

Exercises

Ⅰ. Answer the following questions.

1. List some consumer electronic devices.
2. When was e-mail first put in use?
3. What canTwitter service do?

Ⅱ. Translate the following sentences into Chinese.

1. Internet access provides home users with connectivity to remote computers. As with companies, home users can access information, communicate with other people, and buy products and services with e-commerce. The main benefit now comes from connecting outside of the home.

2. Between person-to-person communications and accessing information are social network applications. Here, the flow of information is driven by the relationships that people declare between each other. One of the most popular social networking sites is Facebook. It lets people update their personal profiles and shares the updates with other people who they have declared to be their friends.

3. Increasingly, consumer electronic devices are networked. For example, some high-end cameras already have a wireless network capability and use it to send photos to a near-by display for viewing. Professional sports photographers can also send their photos to their editors in real-time, first wirelessly to an access point then over the Internet. Devices such as televisions that plug into the wall can use power-line networks to send information throughout the house over the wires that carry electricity.

____

____

____

____

____

Abstract Reading

Algorithms for Enhanced Inter-cell Interference Coordination (eICIC) in LTE HetNets

The success of LTE heterogeneous networks (HetNets) with macrocells and picocells critically depends on efficient spectrum sharing between high-power macros and low-power picos. Two important challenges in this context are: (1) determining the amount of radio resources that macrocells should offer to picocells, and (2) determining the association rules that decide which user equipments (UEs) should associate with picos. In this paper, we develop a novel algorithm to solve these two coupled problems in a joint manner. Our algorithm has provable guarantee, and furthermore, it accounts for network topology, traffic load, and macro-pico interference map. Our solution is standard compliant and can be implemented using the notion of Almost Blank Subframes (ABS) and Cell Selection Bias (CSB) proposed by LTE standards. We also show extensive evaluations using RF plan from a real network and discuss self-optimized networking (SON)-based enhanced inter-cell interference coordination (eICIC) implementation.

Fine Comb: Measuring Microscopic Latency and Loss in the Presence of Reordering

Modern stock trading and cluster applications require microsecond latencies and almost no losses in data centers. This paper introduces an algorithm called FineComb that can obtain fine-grain end-to-end loss and latency measurements between edge routers in these net-

works. Such a mechanism can allow managers to distinguish between latencies and loss singularities caused by servers and those caused by the network. Compared to prior work, such as Lossy Difference Aggregator (LDA), which focused on switch-level latency measurements, the requirement of end-to-end latency measurements introduces the challenge of reordering that occurs commonly in IP networks due to churn. The problem is even more acute in switches across data center networks that employ multipath routing algorithms to exploit the inherent path diversity. Without proper care, a loss estimation algorithm can confound loss and reordering; furthermore, any attempt to aggregate delay estimates in the presence of reordering results in severe errors. FineComb deals with these problems using order-agnostic packet digests and a simple new idea we call stash recovery. Our evaluation demonstrates that FineComb is orders of magnitude more accurate than LDA in loss and delay estimates in the presence of reordering.

Infrastructure-free Content-based Publish/Subscribe

Peer-to-peer (P2P) networks can offer benefits to distributed content-based publish/subscribe data dissemination systems. In particular, since a P2P network's aggregate resources grow as the number of participants increases, scalability can be achieved using no infrastructure other than the participants' own resources. This paper proposes algorithms for supporting content-based publish/subscribe in which subscriptions can specify a range of interest and publications a range of values. The algorithms are built over a distributed hash table abstraction and are completely decentralized. Load balance is addressed by subscription delegation away from overloaded peers and a bottom-up tree search technique that avoids root hotspots. Furthermore, fault tolerance is achieved with a lightweight replication scheme that quickly detects and recovers from faults. Experimental results support the scalability and fault-tolerance properties of the algorithms: for example, doubling the number of subscriptions does not double internal system messages, and even the simultaneous failure of 20% of the peers in the system requires less than 2 min to fully recover.

UNIT 15

DIGITAL COMMUNICATION

Text: Introduction to Digital Communication

Communication has been one of the deepest needs of the human race throughout recorded history. It is essential to forming social unions, to educating the young, and to expressing a myriad of emotions and needs. Good communication is central to a civilized society.

The various communication disciplines in engineering have the purpose of providing technological aids to human communication. One could view the smoke signals and drum rolls of primitive societies as being technological aids to communication, but communication technology as we view it today became important with telegraphy, then telephony, then video, then computer communication, and today the amazing mixture of all of these in inexpensive, small portable devices.

Initially these technologies were developed as separate networks and were viewed as having little in common. As these networks grew, however, the fact that all parts of a given network had to work together, coupled with the fact that different components were developed at different times using different design methodologies, caused an increased focus on the underlying principles and architectural understanding required for continued system evolution.

This need for basic principles was probably best understood at American Telephone and Telegraph (AT&T), where Bell Laboratories was created as the research and development arm of AT&T. The Math Center at Bell Labs became the predominant center for communication research in the world, and held that position until quite recently.

The central core of the principles of communication technology was developed at that center. Perhaps the greatest contribution from the Math Center was the creation of Information Theory by Claude Shannon. For perhaps the first 25 years of its existence, Information Theory was regarded as a beautiful theory but not as a central guide to the architec-



ture and design of communication systems. After that time, however, both the device technology and the engineering understanding of the theory were sufficient to enable system development to follow information theoretic principles.

A number of information theoretic ideas and how they affect communication system design will be explained carefully in later parts. One pair of ideas, however, is central to almost every topic. The first is to view all communication sources, e. g. , speech waveforms, image waveforms, and text files, as being representable by binary sequences. The second is to design communication systems that first convert the source output into a binary sequence and then convert that binary sequence into a form suitable for transmission over particular physical media such as cable, twisted wire pair, optical fiber, or electromagnetic radiation through space.

Digital communication systems, by definition, are communication systems that use such a digital sequence as an interface between the source and the channel input (and similarly between the channel output and final destination).

The idea of converting an analog source output to a binary sequence was quite revolutionary in 1948, and the notion that this should be done before channel processing was even more revolutionary. Today, with digital cameras, digital video, digital voice, etc. , the idea of digitizing any kind of source is commonplace even among the most technophobic. The notion of a binary interface before channel transmission is almost as commonplace. For example, we all refer to the speed of our Internet connection in bits per second.

There are a number of reasons why communication systems now usually contain a binary interface between source and channel (i. e. , why digital communication systems are now standard). These will be explained with the necessary qualifications later, but briefly they are as follows.

(1) Digital hardware has become so cheap, reliable, and miniaturized that digital interfaces are eminently practical.

(2) A standardized binary interface between source and channel simplifies implementation and understanding, since source coding/decoding can be done independently of the channel, and, similarly, channel coding/decoding can be done independently of the source.

(3) A standardized binary interface between source and channel simplifies networking, which now reduces to sending binary sequences through the network.

(4) One of the most important of Shannon's information theoretic results is that if a source can be transmitted over a channel in any way at all, it can be transmitted using a binary interface between source and channel. This is known as the source/channel separation theorem.

In the remainder of this paper, the problems of source coding and decoding and channel coding and decoding are briefly introduced. First, however, the notion of layering in a

communication system is introduced. Standardized interfaces and layering large communication systems such as the Public Switched Telephone Network (PSTN) and the Internet have incredible complexity, made up of an enormous variety of equipment made by different manufacturers at different times following different design principles. Such complex networks need to be based on some simple architectural principles in order to be understood, managed, and maintained. Two such fundamental architectural principles are standardized interfaces and layering.

A standardized interface allows the user or equipment on one side of the interface to ignore all details about the other side of the interface except for certain specified interface characteristics. For example, the binary interface in Figure 15.1 allows the source coding/decoding to be done independently of the channel coding/decoding. The idea of layering in communication systems is to break up communication functions into a string of separate layers. Each layer consists of an input module at the input end of a communication system and a "peer" output module at the other end. The input module at layer i processes the information received from layer $i+1$ and sends the processed information on to layer $i-1$. The peer output module at layer i works in the opposite direction, processing the received information from layer $i-1$ and sending it on to layer i. As an example, an input module might receive a voice waveform from the next higher layer and convert the waveform into a binary data sequence that is passed on to the next lower layer. The output peer module would receive a binary sequence from the next lower layer at the output and convert it back to a speech waveform.

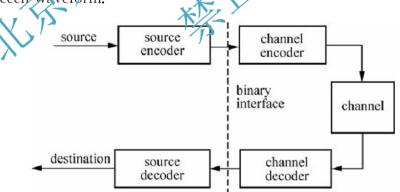


Figure 15.1 Placing a binary interface between source and channel. The source encoder converts the source output to a binary sequence and the channel encoder (often called a modulator) processes the binary sequence for transmission over the channel. The channel decoder (demodulator) recreates the incoming binary sequence (hopefully reliably), and the source decoder recreates the source output.

As another example, a modem consists of an input module (a modulator) and an output module (a demodulator). The modulator receives a binary sequence from the next



higher input layer and generates a corresponding modulated waveform for transmission over a channel. The peer module is the remote demodulator at the other end of the channel. It receives a more or less faithful replica of the transmitted waveform and reconstructs a typically faithful replica of the binary sequence. Similarly, the local demodulator is the peer to a remote modulator (often collocated with the remote demodulator above). Thus a modem is an input module for communication in one direction and an output module for independent communication in the opposite direction. For now, it is enough simply to view the modulator as converting a binary sequence to a waveform, with the peer demodulator converting the waveform back to the binary sequence. As another example, the source coding/decoding layer for a waveform source can be split into three layers. One of the advantages of this layering is that discrete sources are important topics in their own right and correspond to the inner layer. Quantization is also an important topic in its own right. After both of these are understood, waveform sources become quite simple to understand. The channel coding/decoding layer can also be split into several layers, but there are a number of ways to do this which will be discussed later. For example, binary error correction coding/decoding can be used as an outer layer with modulation and demodulation as an inner layer, but it will be seen later that there are a number of advantages in combining these layers into what is called coded modulation. Even here, however, layering is important, but the layers are defined differently for different purposes.

It should be emphasized that layering is much more than simply breaking a system into components. The input and peer output in each layer encapsulate all the lower layers, and all these lower layers can be viewed in aggregate as a communication channel. Similarly, the higher layers can be viewed in aggregate as a simple source and destination.

The above discussion of layering implicitly assumed a point-to-point communication system with one source, one channel, and one destination. Network situations can be considerably more complex. With broadcasting, an input module at one layer may have multiple peer output modules. Similarly, in multiaccess communication a multiplicity of input modules have a single peer output module. It is also possible in network situations for a single module at one level to interface with multiple modules at the next lower layer or the next higher layer. The use of layering is at least as important for networks as it is for point-to-point communications systems.

New Words and Phrases

myriad *adj*. 无数的；种种的
n. 无数，极大数量

telegraphy *n*. 电信；电报学

telephony *n*. 电话(学)；电话制造

methodology *n.* 方法学，方法论

predominant *adj.* 主要的；卓越的；支配的；有力的；有影响的

waveform *n.* （电子）波形

electromagnetic *adj.* 电磁的

technophobic *n.* 恐惧新技术者，恐技术症患者；厌恶计算机者

miniaturize *vt.* 使小型化；使微型化

eminently *adv.* 突出地；显著地

demodulator *n.* 解调器

encapsulate *vt.* 压缩；将……装入胶囊；将……封进内部；概述

aggregate *vi.* 集合；聚集；合计

vt. 集合；聚集；合计

n. 合计；集合体；总计

adj. 聚合的；集合的；合计的

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. portable device
2. speech waveform
3. image waveform
4. text file
5. binary sequence
6. electromagnetic radiation
7. digital communication system
8. binary interface
9. channel separation theorem
10. channel decoding

Ⅱ. Fill in the blanks with missing words from the table below.

capacity	successive	channel	sensitive
compression	restricted	distortion	input
essential	modulation	arbitrarily	expense
waveform	inner	discrete	layers
modulator	analog	probabilistic	unintuitive

1. Frequently the error probability incurred with simple ________ and demodulation techniques is too high. One possible solution is to separate the ________ encoder into two ________: first an error-correcting code, then a simple ________.



2. What Shannon showed was the very ________ fact that more sophisticated coding schemes can achieve ________ low error probability at any data rate below a value known as the channel ________. The channel capacity is a function of the ________ description of the output conditional on each possible ________. Conversely, it is not possible to achieve low error probability at rates above the channel capacity.

3. The output of an ________ source, in the simplest case, is an analog real ________, representing, for example, a speech waveform. The word analog is used to emphasize that the waveform can be arbitrary and is not ________ to taking on amplitudes from some ________ set of values.

4. Discrete source coding is important both for discrete sources, such as text and computer files, and also as an ________ layer for discrete-time analog sequences and fully analog sources. It is ________ to focus on the range of possible outputs from the source rather than any one particular output. It is also important to focus on probabilistic models so as to achieve the best ________ for the most common outputs with less care for very rare outputs.

5. Variable-length source coding is the simplest way to provide good compression for common source outputs at the ________ of rare outputs. The necessity to concatenate ________ variable-length codewords leads to the nonprobabilistic concept of unique decodability.

6. There are some disadvantages to measuring ________ only in a mean-squared sense. For example, efficient speech coders are based on models of human speech. They make use of the fact that human listeners are more ________ to some kinds of reconstruction error than others, so as, for example, to permit larger errors when the signal is loud than when it is soft.

Ⅲ. Translate the following sentences into Chinese.

1. The first is to understand analog data compression, i. e., the compression of sources such as voice for which the output is an arbitrarily varying real- or complex-valued function of time; we denote such functions as waveforms. The second is to begin studying the waveforms that are typically transmitted at the input and received at the output of communication channels. The same set of mathematical tools is required for the understanding and representation of both source and channel waveforms.

____

____

____

____

____

2. A vector space is essentially a collection of objects (such as the collection of real n-tuples) along with a set of rules for manipulating those objects. There is a set of axioms describing precisely how these objects and rules work. Any properties that follow from those axioms must then apply to any vector space, i. e., any set of objects satisfying those axioms.

____

____

____

____

____

3. Digital modulation (or channel encoding) is the process of converting an input sequence of bits into a waveform suitable for transmission over a communication channel. Demodulation (channel decoding) is the corresponding process at the receiver of converting the received waveform into a (perhaps noisy) replica of the input bit sequence.

____

____

____

____

____

Reading: Detection, Coding, and Decoding

This paper uses that characterization to retrieve the signal from the noise-corrupted received waveform. As one might guess, this is not possible without occasional errors when the noise is unusually large. The objective is to retrieve the data while minimizing the effect of these errors. This process of retrieving data from a noise-corrupted version is known as detection.

Detection, decision making, hypothesis testing, and decoding are synonyms. The word detection refers to the effort to detect whether some phenomenon is present or not on the basis of observations. For example, a radar system uses observations to detect whether or not a target is present; a quality control system attempts to detect whether a unit is defective; a medical test detects whether a given disease is present. The meaning of detection has been extended in the digital communication field from a yes/no decision to a decision at the receiver between a finite set of possible transmitted signals. Such a decision between a set of possible transmitted signals is also called decoding, but here the possible set is usual-



ly regarded as the set of codewords in a code rather than the set of signals in a signal set. Decision making is, again, the process of deciding between a number of mutually exclusive alternatives. Hypothesis testing is the same, but here the mutually exclusive alternatives are called hypotheses. We use the word hypotheses for the possible choices in what follows, since the word conjures up the appropriate intuitive image of making a choice between a set of alternatives, where only one alternative is correct and there is a possibility of erroneous choice.

These problems will be studied initially in a purely probabilistic setting. That is, there is a probability model within which each hypothesis is an event. These events are mutually exclusive and collectively exhaustive; i. e. , the sample outcome of the experiment lies in one and only one of these events, which means that in each performance of the experiment, one and only one hypothesis is correct. Assume there are M hypotheses, labeled $a_0, \cdots, a_{M-1}$. The sample outcome of the experiment will be one of these M events, and this defines a random symbol U which, for each m, takes the value a_m when event occurs. The marginal probability $p_U(a_m)$ of hypothesis a_m is denoted by p_m and is called the a-priori probability of a_m. There is also a random variable (rv) V, called the observation. A sample value v of V is observed, and on the basis of that observation the detector selects one of the possible M hypotheses. The observation could equally well be a complex random variable, a random vector, a random process, or a random symbol; these generalizations are discussed in what follows.

Before discussing how to make decisions, it is important to understand when and why decisions must be made. For a binary example, assume that the conditional probability of hypothesis a_0 given the observation is 2/3 and that of hypothesis a_1 is 1/3. Simply deciding on hypothesis a_0 and forgetting about the probabilities throws away the information about the probability that the decision is correct. However, actual decisions sometimes must be made. In a communication system, the user usually wants to receive the message (even partly garbled) rather than a set of probabilities. In a control system, the controls must occasionally take action. Similarly, managers must occasionally choose between courses of action, between products, and between people to hire. In a sense, it is by making decisions that we return from the world of mathematical probability models to the world being modeled.

There are a number of possible criteria to use in making decisions. Initially assume that the criterion is to maximize the probability of correct choice. That is, when the experiment is performed, the resulting experimental outcome maps into both a sample value a_m for U and a sample value v for V. The decision maker observes v (but not a_m) and maps v into a decision $\tilde{u}(v)$. The decision is correct if $\tilde{u}(v) = a_m$. In principle, maximizing the proba-

bility of correct choice is almost trivially simple. Given v, calculate $p_{U|V}(a_m \mid v)$ for each possible hypothesis a_m. This is the probability that a_m is the correct hypothesis conditional on v. Thus the rule for maximizing the probability of being correct is to choose $\tilde{u}(v)$ to be that a_m for which $p_{U|V}(a_m \mid v)$ is maximized. For each possible observation v, this is denoted by

$$\tilde{u}(v) = \arg\max_m p_{U|V}(a_m \mid v) \quad \text{(MAP rule)} \tag{15-1}$$

where arg max means the argument m that maximizes the function. If the maximum is not unique, the probability of being correct is the same no matter which maximizing m is chosen, so, to be explicit, the smallest such m will be chosen. Since the rule (15-1) applies to each possible sample output v of the random variable V, (15-1) also defines the selected hypothesis as a random symbol $\tilde{U}(V)$. The conditional probability $p_{U|V}$ is called an a-posteriori probability. This is in contrast to the a-priori probability p_U of the hypothesis before the observation of V. The decision rule in (15-1) is thus called the maximum a-posteriori probability (MAP) rule.

An important consequence of (15-1) is that the MAP rule depends only on the conditional probability $p_{U|V}$ and thus is completely determined by the joint distribution of U and V. Everything else in the probability space is irrelevant to making a MAP decision.

When distinguishing between different decision rules, the MAP decision rule in (15-1) will be denoted by $\tilde{u}_{\text{MAP}}(v)$. Since the MAP rule maximizes the probability of correct decision for each sample value v, it also maximizes the probability of correct decision averaged over all v. To see this analytically, let $\tilde{u}_D(v)$ be an arbitrary decision rule. Since $\tilde{u}_{\text{MAP}}$ maximizes $p_{U|V}(a_m \mid v)$ over m,

$$p_{U|V}(\tilde{u}_{\text{MAP}}(v) \mid v) - p_{U|V}(\tilde{u}_D(v) \mid v) \geqslant 0 \text{ for each rule } D \text{ and observation } v \tag{15-2}$$

Taking the expected value of the first term on the left over the observation V, we get the probability of correct decision using the MAP decision rule. The expected value of the second term on the left for any given I is the probability of correct decision using that rule. Thus, taking the expected value of (15-2) over V shows that the MAP rule maximizes the probability of correct decision over the observation space. The above results are very simple, but also important and fundamental. They are summarized in the following theorem. Before discussing the implications and use of the MAP rule, the above assumptions are reviewed. First, a probability model was assumed in which all probabilities are known, and in which, for each performance of the experiment, one and only one hypothesis is correct. This conforms very well to the communication model in which a transmitter sends one of a set of possible signals and the receiver, given signal plus noise, makes a decision on the signal actually sent. It does not always conform well to a scientific experiment attempting



to verify the existence of some new phenomenon; in such situations, there is often no sensible way to model a-priori probabilities. Detection in the absence of known a-priori probabilities is discussed in later parts.

The next assumption was that maximizing the probability of correct decision is an appropriate decision criterion. In many situations, the cost of a wrong decision is highly asymmetric. For example, when testing for a treatable but deadly disease, making an error when the disease is present is far more costly than making an error when the disease is not present. It is easy to extend the theory to account for relative costs of errors. With the present assumptions, the detection problem can be stated concisely in the following probabilistic terms. There is an underlying sample space, a probability measure, and two rvs U and V of interest. The corresponding experiment is performed, an observer sees the sample value v of rv V, but does not observe anything else, particularly not the sample value of U, say a_m. The observer uses a detection rule, $\tilde{u}(v)$, which is a function mapping each possible value of v to a possible value of U. If $\tilde{u}(v) = a_m$, the detection is correct; otherwise an error has been made. The above MAP rule maximizes the probability of correct detection conditional on each v and also maximizes the unconditional probability of correct detection. Obviously, the observer must know the conditional probability assignment $p_{U|V}$ in order to use the MAP rule.

New Words and Phrases

hypothesis *n.* 假设
conjure *vt.* 提出，想象
intuitive *adj.* 直觉的；凭直觉获知的
erroneous *adj.* 错误的；不正确的
probabilistic *adj.* 概率的
garble *vt.* 断章取义；歪曲；混淆
　　n. 断章取义；混淆；篡改
criteria *n.* 标准
conform *vi.* 符合；遵照；适应环境
　　vt. 使遵守；使一致；使顺从
asymmetric *adj.* 不对称的；非对称的
a-posteriori *adj.* 后验的

Exercises

Ⅰ. Answer the following questions.

1. What does "detection" refer to in this paper?
2. What is the MAP rule?
3. Which decision criterion is appropriate to decoding?

Ⅱ. Translate the following sentences into Chinese.

1. Detection, decision making, hypothesis testing, and decoding are synonyms. The word detection refers to the effort to detect whether some phenomenon is present or not on the basis of observations. For example, a radar system uses observations to detect whether or not a target is present; a quality control system attempts to detect whether a unit is defective; a medical test detects whether a given disease is present.

2. As one might guess, this is not possible without occasional errors when the noise is unusually large. The objective is to retrieve the data while minimizing the effect of these errors. This process of retrieving data from a noise-corrupted version is known as detection.

3. The meaning of detection has been extended in the digital communication field from a yes/no decision to a decision at the receiver between a finite set of possible transmitted signals. Such a decision between a set of possible transmitted signals is also called decoding, but here the possible set is usually regarded as the set of codewords in a code rather than the set of signals in a signal set.



Abstract Reading

Spatially-coupled LDPC Codes for Decode-and-forward Relaying of Two Correlated Sources over the BEC

We present a decode-and-forward transmission scheme based on spatially-coupled low-density parity-check (SC-LDPC) codes for a network consisting of two (possibly correlated) sources, one relay, and one destination. The links between the nodes are modeled as binary erasure channels. Joint source-channel coding with joint channel decoding is used to exploit the correlation. The relay performs network coding. We derive analytical bounds on the achievable rates for the binary erasure time-division multiple-access relay channel with correlated sources. We then design bilayer SC-LDPC codes and analyze their asymptotic performance for this scenario. We prove analytically that the proposed coding scheme achieves the theoretical limit for symmetric channel conditions and uncorrelated sources. Using density evolution, we furthermore demonstrate that our scheme approaches the theoretical limit also for non-symmetric channel conditions and when the sources are correlated, and we observe the threshold saturation effect that is typical for spatially-coupled systems. Finally, we give simulation results for large block lengths, which validate the DE analysis.

A Tractable Approach to Coverage and Rate in Cellular Networks

Cellular networks are usually modeled by placing the base stations on a grid, with mobile users either randomly scattered or placed deterministically. These models have been used extensively but suffer from being both highly idealized and not very tractable, so complex system-level simulations are used to evaluate coverage/outage probability and rate. More tractable models have long been desirable. We develop new general models for the multi-cell signal-to-interference-plus-noise ratio (SINR) using stochastic geometry. Under very general assumptions, the resulting expressions for the downlink SINR CCDF (equivalent to the coverage probability) involve quickly computable integrals, and in some practical special cases can be simplified to common integrals (e. g., the Q-function) or even to simple closed-form expressions. We also derive the mean rate, and then the coverage gain (and mean rate loss) from static frequency reuse. We compare our coverage predictions to

the grid model and an actual base station deployment, and observe that the proposed model is pessimistic (a lower bound on coverage) whereas the grid model is optimistic, and that both are about equally accurate. In addition to being more tractable, the proposed model may better capture the increasingly opportunistic and dense placement of base stations in future networks.

A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction

This paper discusses the effects of frequency offset on the performance of orthogonal frequency division multiplexing (OFDM) digital communications. The main problem with frequency offset is that it introduces interference among the multiplicity of carriers in the OFDM signal. It is shown, and confirmed by simulation, that to maintain signal-to-interference ratios of 20 dB or greater for the OFDM carriers, offset is limited to 4% or less of the intercarrier spacing. Next, the paper describes a technique to estimate frequency offset using a repeated data symbol. A maximum likelihood estimation (MLE) algorithm is derived and its performance computed and compared with simulation results. Since the intercarrier interference energy and signal energy both contribute coherently to the estimate, the algorithm generates extremely accurate estimates even when the offset is far too great to demodulate the data values. Also, the estimation error depends only on total symbol energy so it is insensitive to channel spreading and frequency selective fading. A strategy is described for initial acquisition in the event of uncertainty in the initial offset that exceeds 1/2 the carrier spacing, the limit of the MLE algorithm.

UNIT 16

DIGITAL SIGNAL PROCESSING

Text: The Breadth and Depth of DSP I

Digital Signal Processing is one of the most powerful technologies that will shape science and engineering in the twenty-first century. Revolutionary changes have already been made in a broad range of fields: communications, medical imaging, radar & sonar, high fidelity music reproduction, and oil prospecting, to name just a few. Each of these areas has developed a deep DSP technology, with its own algorithms, mathematics, and specialized techniques. This combination of breadth and depth makes it impossible for any one individual to master all of the DSP technology that has been developed. DSP education involves two tasks: learning general concepts that apply to the field as a whole, and learning specialized techniques for your particular area of interest. This chapter starts our journey into the world of Digital Signal Processing by describing the dramatic effect that DSP has made in several diverse fields. The revolution has begun.

The Roots of DSP

Digital Signal Processing is distinguished from other areas in computer science by the unique type of data it uses: signals. In most cases, these signals originate as sensory data from the real world: seismic vibrations, visual images, sound waves, etc.. DSP is the mathematics, the algorithms, and the techniques used to manipulate these signals after they have been converted into a digital form. This includes a wide variety of goals, such as: enhancement of visual images, recognition and generation of speech, compression of data for storage and transmission, etc.. Suppose we attach an analog-to-digital converter to a computer and use it to acquire a chunk of real world data. DSP answers the question:

What next?

The roots of DSP are in the 1960s and 1970s when digital computers first became available. Computers were expensive during this era, and DSP was limited to only a few critical applications. Pioneering efforts were made in four key areas: radar & sonar, where national security was at risk; oil exploration, where large amounts of money could be made; space exploration, where the data are irreplaceable; and medical imaging, where lives could be saved. The personal computer revolution of the 1980s and 1990s caused DSP to explode with new applications. Rather than being motivated by military and government needs, DSP was suddenly driven by the commercial marketplace. Anyone who thought they could make money in the rapidly expanding field was suddenly a DSP vendor. DSP reached the public in such products as: mobile telephones, compact disc players, and electronic voice mail.

This technological revolution occurred from the top-down. In the early 1980s, DSP was taught as agraduate level course in electrical engineering. A decade later, DSP had become a standard part of the undergraduate curriculum. Today, DSP is a basic skill needed by scientists and engineers in many fields. As an analogy, DSP can be compared to a previous technological revolution: electronics. While still the realm of electrical engineering, nearly every scientist and engineer has some background in basic circuit design. Without it, they would be lost in the technological world. DSP has the same future.

This recent history is more than a curiosity; it has a tremendous impact onyour ability to learn and use DSP. Suppose you encounter a DSP problem, and turn to textbooks or other publications to find a solution. What you will typically find is page after page of equations, obscure mathematical symbols, and unfamiliar terminology. It's a nightmare! Much of the DSP literature is baffling even to those experienced in the field. It's not that there is anything wrong with this material, it is just intended for a very specialized audience. State-of-the-art researchers need this kind of detailed mathematics to understand the theoretical implications of the work.

A basic premise of this book is that most practical DSP techniques can be learned and used without the traditional barriers of detailed mathematics and theory. The Scientist and Engineer's Guide to Digital Signal Processing is written for those who want to use DSP as a tool, not a new career.

The remainder of this paper illustrates areas where DSP has produced revolutionary changes. As you go through each application, notice that DSP is very interdisciplinary, relying on the technical work in many adjacent fields. The borders between DSP and other technical disciplines are not sharp and well defined, but rather fuzzy and overlapping. If you want to specialize in DSP, these are the allied areas you will also need to study.



Telecommunications

Telecommunications is about transferring information from one location to another. This includes many forms of information: telephone conversations, television signals, computer files, and other types of data. To transfer the information, you need achannel between the two locations. This may be a wire pair, radio signal, optical fiber, etc.. Telecommunications companies receive payment for transferring their customer's information, while they must pay to establish and maintain the channel. The financial bottom line is simple: the more information they can pass through a single channel, the more money they make. DSP has revolutionized the telecommunications industry in many areas: signaling tone generation and detection, frequency band shifting, filtering to remove power line hum, etc.. Three specific examples from the telephone network will be discussed here: multiplexing, compression, and echo control.

Multiplexing

There are approximately one billion telephones in the world. At the press of a few buttons, switching networks allow any one of these to be connected to any other in only a few seconds. The immensity of this task is mind boggling! Until the 1960s, a connection between two telephones required passing the analog voice signals through mechanical switches and amplifiers. One connection required one pair of wires. In comparison, DSP converts audio signals into a stream of serial digital data. Since bits can be easily intertwined and later separated, many telephone conversations can be transmitted on a single channel. For example, a telephone standard known as the T-carrier system can simultaneously transmit 24 voice signals. Each voice signal is sampled 8 000 times per second using an 8 bit companded (logarithmic compressed) analog-to-digital conversion. This results in each voice signal being represented as 64 000 bps, and all 24 channels being contained in 1.544 Mbps. This signal can be transmitted about 6 000 feet using ordinary telephone lines of 22 gauge copper wire, a typical interconnection distance. The financial advantage of digital transmission is enormous. Wire and analog switches are expensive; digital logic gates are cheap.

Compression

When a voice signal is digitized at 8 000 samples/sec, most of the digital information is redundant. That is, the information carried by any one sample is largely duplicated by the

neighboring samples. Dozens of DSP algorithms have been developed to convert digitized voice signals into data streams that require fewer bps. These are called data compression algorithms. Matching uncompression algorithms are used to restore the signal to its original form. These algorithms vary in the amount of compression achieved and the resulting sound quality. In general, reducing the data rate from 64 kbps to 32 kbps results in no loss of sound quality. When compressed to a data rate of 8 kbps, the sound is noticeably affected, but still usable for long distance telephone networks. The highest achievable compression is about 2 kbps, resulting in sound that is highly distorted, but usable for some applications such as military and undersea communications.

Echo Control

Echoes are a serious problem in long distance telephone connections. When you speak into a telephone, a signal representing your voice travels to the connecting receiver, where a portion of it returns as an echo. If the connection is within a few hundred miles, the elapsed time for receiving the echo is only a few milliseconds. The human ear is accustomed to hearing echoes with these small time delays, and the connection sounds quite normal. As the distance becomes larger, the echo becomes increasingly noticeable and irritating. The delay can be several hundred milliseconds for intercontinental communications, and is particularly objectionable.

Digital Signal Processing attacks this type of problem by measuring the returned signal and generating an appropriateantisignal to cancel the offending echo. This same technique allows speakerphone users to hear and speak at the same time without fighting audio feedback (squealing). It can also be used to reduce environmental noise by canceling it with digitally generated antinoise.

New Words and Phrases

sonar *n.* 声呐；声波定位仪
sensory *adj.* 感觉的；知觉的；传递感觉的
seismic *adj.* 地震的
vibration *n.* 振动
chunk *n.* 大块
obscure *adj.* 昏暗的，朦胧的；晦涩的，不清楚的；隐蔽的；不著名的，无名的
vt. 使……模糊不清，掩盖；隐藏；使难理解
baffle *vt.* 使……困惑；使……受挫折；用挡板控制
interdisciplinary *adj.* 各学科间的
fuzzy *adj.* 模糊的；失真的
overlap *vt.* 与……重叠；与……同时发生
filter *n.* 滤波器
multiplexing *n.* 多路复用



echo *n.* 回音；效仿
immensity *n.* 巨大；无限
boggle *vi.* 犹豫，退缩；惊恐
vt. 搞糟，弄坏；使……惊奇；使……困惑
n. 犹豫，退缩；惊奇
intertwine *vt.* 缠绕；纠缠
compand *n.* 压缩扩展
logarithmic *adj.* 对数的
megabits *n.* 兆位
gauge *n.* 计量器；标准尺寸；容量规格
vt. 测量；估计；给……定规格
duplicate *vt.* 复制；使加倍
n. 副本；复制品
distort *vt.* 扭曲；使失真；曲解
elapse *vi.* 消逝；时间过去；
n. 流逝；时间的过去

Exercises

Ⅰ. Translate the following phrases into Chinese.

1. Digital Signal Processing
2. medical imaging
3. seismic vibration
4. circuit design
5. financial bottom line
6. echo control
7. analog voice signals
8. analog-to-digital conversion
9. data streams
10. elapsed time

Ⅱ. Fill in the blanks with missing words from the table below.

application	filters	reactions	continuously
functionality	systematic	linear	fundamentally
diagnostic	sensors	classification	implement
analog	hypotheses	compliance	capabilities
embedded	sequence	processing	components

1. The primary traits of ________ signal processing systems that distinguish them from general purpose computer systems are their predictable ________ to real-time stimuli from the environment, their form-and cost-optimized design, and their ________ with re-



quired or specified modes of response behavior and ________.

2. An embedded system usually consists of hardware ________ such as memories, ________-specific ICs (ASICs), processors, DSPs, buses, analog-digital interfaces, and also software components that provide control, ________, and application-specific ________ required of it. In addition, they often contain electromechanical (EM) components such as ________ and transducers and operate in harsh environmental conditions.

3. Digital signal processing methods ________ require that signals are quantized at discrete time instances and represented as a ________ of words consisting of 1's and 0's. In nature, signals are usually nonquantized and ________ varied with time. Natural signals such as air pressure waves as a result of speech are converted by a transducer to a proportional ________ electrical signal.

4. Digital ________ are widely used in processing digital signals of many diverse applications, including speech processing and data communications, image and video ________, sonar, radar, seismic and oil exploration, and consumer electronics. One class of digital filters, the ________ shift-invariant (LSI) type, are the most frequently used because they are simple to analyze, design, and ________.

5. Detection and classification arise in signal processing problems whenever a decision is to be made among a finite number of ________ concerning an observed waveform. Signal detection algorithms decide whether the waveform consists of "noise alone" or "signal masked by noise". Signal classification algorithms decide whether a detected signal belongs to one or another of prespecified classes of signals. The objective of signal detection and ________ theory is to specify ________ strategies for designing algorithms which minimize the average number of decision errors.

Ⅲ. Translate the following sentences into Chinese.

1. Images are signals with special characteristics. First, they are a measure of a parameter over space (distance), while most signals are a measure of a parameter over time. Second, they contain a great deal of information. For example, more than 10 megabytes can be required to store one second of television video. This is more than a thousand times greater than for a similar length voice signal. Third, the final judge of quality is often a subjective human evaluation, rather than an objective criterion. These special characteristics have made image processing a distinct subgroup within DSP.

____

____

____

____



2. Sometimes, you just have to make the most out of a bad picture. This is frequently the case with images taken from unmanned satellites and space exploration vehicles. No one is going to send a repairman to Mars just to tweak the knobs on a camera! DSP can improve the quality of images taken under extremely unfavorable conditions in several ways: brightness and contrast adjustment, edge detection, noise reduction, focus adjustment, motion blur reduction, etc.. Images that have spatial distortion, such as encountered when a flat image is taken of a spherical planet, can also bewarped into a correct representation. Many individual images can also be combined into a single database, allowing the information to be displayed in unique ways. For example, a video sequence simulating an aerial flight over the surface of a distant planet.

Reading: The Breadth and Depth of DSP II

Audio Processing

The two principal human senses are vision and hearing. Correspondingly, much of DSP is related to image and audio processing. People listen to both music and speech. DSP has made revolutionary changes in both these areas.

Music

The path leading from the musician's microphone to the audiophile's speaker is remarkably long. Digital data representation is important to prevent the degradation commonly associated with analog storage and manipulation. This is very familiar to anyone who has compared the musical quality of cassette tapes with compact disks. In a typical

scenario, a musical piece is recorded in a sound studio on multiple channels or tracks. In some cases, this even involves recording individual instruments and singers separately. This is done to give the sound engineer greater flexibility in creating the final product. The complex process of combining the individual tracks into a final product is called mix down. DSP can provide several important functions during mix down, including: filtering, signal addition and subtraction, signal editing, etc..

One of the most interesting DSP applications in music preparation is artificial reverberation. If the individual channels are simply added together, the resulting piece sounds frail and diluted, much as if the musicians were playing outdoors. This is because listeners are greatly influenced by the echo or reverberation content of the music, which is usually minimized in the sound studio. DSP allows artificial echoes and reverberation to be added during mix down to simulate various ideal listening environments. Echoes with delays of a few hundred milliseconds give the impression of cathedral-like locations. Adding echoes with delays of 10-20 milliseconds provide the perception of more modest size listening rooms.

Speech Generation

Speech generation and recognition are used to communicate between humans and machines. Rather than using your hands and eyes, you use your mouth and ears. This is very convenient when your hands and eyes should be doing something else, such as: driving a car, performing surgery, or (unfortunately) firing your weapons at the enemy. Two approaches are used for computer generated speech: digital recording and vocal tract simulation. In digital recording, the voice of a human speaker is digitized and stored, usually in a compressed form. During playback, the stored data are uncompressed and converted back into an analog signal. An entire hour of recorded speech requires only about three megabytes of storage, well within the capabilities of even small computer systems. This is the most common method of digital speech generation used today.

Vocal tract simulators are more complicated, trying to mimic the physical mechanisms by which humans create speech. The human vocal tract is an acoustic cavity with resonant frequencies determined by the size and shape of the chambers. Sound originates in the vocal tract in one of two basic ways, called voiced and fricative sounds. With voiced sounds, vocal cord vibration produces near periodic pulses of air into the vocal cavities. In comparison, fricative sounds originate from the noisy air turbulence at narrow constrictions, such as the teeth and lips. Vocal tract simulators operate by generating digital signals that resemble these two types of excitation. The characteristics of the resonate chamber are simulated by passing the excitation signal through a digital filter with similar resonances. This approach was used in one of the very early DSP success stories, the Speak & Spell, a wide-



ly sold electronic learning aid for children.

Speech Recognition

The automated recognition of human speech is immensely more difficult than speech generation. Speech recognition is a classic example of things that the human brain does well, but digital computers do poorly. Digital computers can store and recall vast amounts of data, perform mathematical calculations at blazing speeds, and do repetitive tasks without becoming bored or inefficient. Unfortunately, present day computers perform very poorly when faced with raw sensory data. Teaching a computer to send you a monthly electric bill is easy. Teaching the same computer to understand your voice is a major undertaking.

Digital Signal Processing generally approaches the problem of voice recognition in two steps: feature extraction followed by feature matching. Each word in the incoming audio signal is isolated and then analyzed to identify the type of excitation and resonate frequencies. These parameters are then compared with previous examples of spoken words to identify the closest match. Often, these systems are limited to only a few hundred words; can only accept speech with distinct pauses between words; and must be retrained for each individual speaker. While this is adequate for many commercial applications, these limitations are humbling when compared to the abilities of human hearing. There is a great deal of work to be done in this area, with tremendous financial rewards for those that produce successful commercial products.

Echo Location

A common method of obtaining information about a remote object is to bounce a wave off of it. For example, radar operates by transmitting pulses of radio waves, and examining the received signal for echoes from aircraft. In sonar, sound waves are transmitted through the water to detect submarines and other submerged objects. Geophysicists have long probed the earth by setting off explosions and listening for the echoes from deeply buried layers of rock. While these applications have a common thread, each has its own specific problems and needs. Digital Signal Processing has produced revolutionary changes in all three areas.

Radar

Radar is an acronym for *Radio Detection And Ranging*. In the simplest radar system, a radio transmitter produces a pulse of radio frequency energy a few microseconds long.

This pulse is fed into a highly directional antenna, where the resulting radio wave propagates away at the speed of light. Aircraft in the path of this wave will reflect a small portion of the energy back toward a receiving antenna, situated near the transmission site. The distance to the object is calculated from the elapsed time between the transmitted pulse and the received echo. The direction to the object is found more simply; you know where you pointed the directional antenna when the echo was received.

The operating range of a radar system is determined by two parameters: how much energy is in the initial pulse, and the noise level of the radio receiver. Unfortunately, increasing the energy in the pulse usually requires making the pulse longer. In turn, the longer pulse reduces the accuracy and precision of the elapsed time measurement. This results in a conflict between two important parameters: the ability to detect objects at long range, and the ability to accurately determine an object's distance.

DSP has revolutionized radar in three areas, all of which relate to this basic problem. First, DSP can compress the pulse after it is received, providing better distance determination without reducing the operating range. Second, DSP can filter the received signal to decrease the noise. This increases the range, without degrading the distance determination. Third, DSP enables the rapid selection and generation of different pulse shapes and lengths. Among other things, this allows the pulse to be optimized for a particular detection problem. Now the impressive part: much of this is done at a sampling rate comparable to the radio frequency used, as high as several hundred megahertz! When it comes to radar, DSP is as much about high-speed hardware design as it is about algorithms.

New Words and Phrases

degradation *n.* 退化；降格，降级；堕落
reverberation *n.* 反响；回响；反射；反射物
frail *adj.* 脆弱的；虚弱的
dilute *adj.* 稀释的；淡的
  vt. 稀释；冲淡；削弱
simulate *vt.* 仿真
cathedral *n.* 大教堂
megabyte *n.* 兆字节
vocal *adj.* 歌唱的；声音的，有声的
  n. 声乐作品；元音
acoustic *adj.* 声学的；音响的；听觉的
cavity *n.* 腔；洞，凹处
resonant *adj.* 共振的；共鸣的
periodic *adj.* 周期的；定期的
excitation *n.* 激发，刺激；激励
resonate *vt.* 共振
resonance *n.* 共振
blazing *adj.* 燃烧的；强烈的；闪耀的
extraction *n.* 取出；抽出；拔出；抽出物
humble *adj.* 谦逊的；简陋的；(级别或地位)低下的；不大的
bounce *n.* 跳；弹力；活力
  vt. 弹跳；使弹起



acronym *n.* 首字母缩略词
pulse *n.* 脉冲；脉搏
vt. 使跳动
antenna *n.* 天线
propagate *vt.* 传播；传送
megahertz *n.* 兆赫

Exercises

Ⅰ. Answer the following questions.

1. What are the two approaches used for computer generated speech?

2. What are the two steps that Digital Signal Processing generally approaches the problem of voice recognition?

3. What's the common method of obtaining information about a remote object?

Ⅱ. Translate the following sentences into Chinese.

1. The path leading from the musician's microphone to the audiophile's speaker is remarkably long. Digital data representation is important to prevent the degradation commonly associated with analog storage and manipulation. This is very familiar to anyone who has compared the musical quality of cassette tapes with compact disks.

____
____
____
____
____

2. The automated recognition of human speech is immensely more difficult than speech generation. Speech recognition is a classic example of things that the human brain does well, but digital computers do poorly. Digital computers can store and recall vast amounts of data, perform mathematical calculations at blazing speeds, and do repetitive tasks without becoming bored or inefficient.

____
____
____
____
____

3. A common method of obtaining information about a remote object is to bounce a wave off of it. For example, radar operates by transmitting pulses of radio waves, and ex-

amining the received signal for echoes from aircraft. In sonar, sound waves are transmitted through the water to detect submarines and other submerged objects.

____

____

____

____

____

Abstract Reading

Source Separation Using Regularized NMF with MMSE Estimates under GMM Priors with Online Learning for the Uncertainties

We propose a new method to incorporate priors on the solution of nonnegative matrix factorization (NMF). The NMF solution is guided to follow the minimum mean square error (MMSE) estimates of the weight combinations under a Gaussian mixture model (GMM) prior. The proposed algorithm can be used for denoising or single-channel source separation (SCSS) applications. NMF is used in SCSS in two main stages, the training stage and the separation stage. In the training stage, NMF is used to decompose the training data spectrogram for each source into a multiplication of a trained basis and gains matrices. In the separation stage, the mixed signal spectrogram is decomposed as a weighted linear combination of the trained basis matrices for the source signals. In this work, to improve the separation performance of NMF, the trained gains matrices are used to guide the solution of the NMF weights during the separation stage. The trained gains matrix is used to train a prior GMM that captures the statistics of the valid weight combinations that the columns of the basis matrix can receive for a given source signal. In the separation stage, the prior GMMs are used to guide the NMF solution of the gains/weights matrices using MMSE estimation. The NMF decomposition weights matrix is treated as a distorted image by a distortion operator, which is learned directly from the observed signals. The MMSE estimate of the weights matrix under the trained GMM prior and log-normal distribution for the distortion is then found to improve the NMF decomposition results. The MMSE estimate is embedded within the optimization objective to form a novel regularized NMF cost function. The corresponding update rules for the new objectives are derived in this paper. The proposed MMSE estimates based regularization avoids the problem of computing the



hyper-parameters and the regularization parameters. MMSE also provides a better estimate for the valid gains matrix. Experimental results show that the proposed regularized NMF algorithm improves the source separation performance compared with using NMF without a prior or with other prior models.

Robust Feature Extraction Based on an Asymmetric Level-dependent Auditory Filterbank and a Subband Spectrum Enhancement Technique

In this paper we introduce a robust feature extractor, dubbed as robust compressive gammachirp filterbank cepstral coefficients (RCGCC), based on an asymmetric and level-dependent compressive gammachirp filterbank and a sigmoid shape weighting rule for the enhancement of speech spectra in the auditory domain. The goal of this work is to improve the robustness of speech recognition systems in additive noise and real-time reverberant environments. As a post processing scheme we employ a short-time feature normalization technique called short-time cepstral mean and scale normalization (STCMSN), which, by adjusting the scale and mean of cepstral features, reduces the difference of cepstra between the training and test environments. For performance evaluation, in the context of speech recognition, of the proposed feature extractor we use the standard noisy AURORA-2 connected digit corpus, the meeting recorder digits (MRDs) subset of the AURORA-5 corpus, and the AURORA-4 LVCSR corpus, which represent additive noise, reverberant acoustic conditions and additive noise as well as different microphone channel conditions, respectively. The ETSI advanced front-end (ETSI-AFE), the recently proposed power normalized cepstral coefficients (PNCC), conventional MFCC and PLP features are used for comparison purposes. Experimental speech recognition results demonstrate that the proposed method is robust against both additive and reverberant environments. The proposed method provides comparable results to that of the ETSI-AFE and PNCC on the AURORA-2 as well as AURORA-4 corpora and provides considerable improvements with respect to the other feature extractors on the AURORA-5 corpus.

Directional Acoustic Source Orientation Estimation Using Only Two Microphones

A simple physical model consisting of a point source displaced from its center of rotation, in combination with a directivity model that includes backwards emitted energy, is considered for the problem of estimating the orientation of a directional acoustic source. Such a problem arises, for instance, in voice-commanded devices in a smart room and is usually tackled with a large or distributed microphone array. We show, however, that when

the time difference of arrival is also taken into account, a small array of only two microphones is sufficiently robust against unaccounted factors such as microphone directivity variation and mild reverberation. This is shown by comparing predicted and measured values of binaural cues, and by using them and pairwise frame energies as inputs for an artificial neural network (ANN) in order to estimate source orientation.

